LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

FACTORES DE PREDICCIÓN DE LA APARICIÓN DE PERSONAS MAYORES REPORTADAS COMO DESAPARECIDAS, A PARTIR DE MODELOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO

PREDICTORS OF FINDING OLDER ADULTS REPORTED MISSING BASED ON SUPERVISED MACHINE LEARNING MODELS

Adriana Lucía Ruiz Rizzo, alruizr@libertadores.edu.co
John González Veloza, jjgonzalezv02@libertadores.edu.co

RESUMEN

La desaparición de personas es un fenómeno enigmático y frecuente que puede traer consecuencias negativas para la persona desaparecida y sus familiares, pero también para la sociedad en general. Por eso, es de vital importancia poder entender mejor el fenómeno de la desaparición y así encontrar estrategias para que se resuelva de la manera más favorable para todos. Los adultos mayores pueden ser particularmente vulnerables al fenómeno de la desaparición debido a los cambios cognitivos y la elevada vulnerabilidad a la demencia que ocurren con el envejecimiento. Por eso, en el presente estudio se buscó identificar los factores individuales y del entorno que pueden pronosticar la aparición de una persona mayor con la ayuda de modelos de aprendizaje automático supervisado. A partir del análisis de datos abiertos sobre las desapariciones en Colombia desde 1930 a junio de 2021 (*n* = 7855) se entrenaron diversos modelos de clasificación utilizando aprendizaje automático supervisado para pronosticar la aparición de personas reportadas como desaparecidas. Los modelos de clasificación con el mejor desempeño en los datos de prueba fueron modelos basados en árboles de decisión, en especial el modelo de máquina de refuerzo de gradiente ligero (*Light Gradient Boosting Machine*), el cual mostró 71% de exactitud en la clasificación (*i.e.*, 8% más que la de un modelo base construido con la media de la duración del reporte de desaparición). Las categorías que más contribuyeron a la predicción fueron la fecha

y el lugar de la desaparición, así como también la edad y el sexo de la persona desaparecida. Los presentes resultados pueden ayudar a entender mejor el fenómeno de la desaparición en personas mayores al tiempo que pueden tener implicaciones prácticas para los casos reales.
Palabras clave: *Adultos mayores; Aprendizaje automático; Clasificación; Desapariciones; Envejecimiento*

ABSTRACT

Person missingness is an enigmatic and frequent phenomenon that can bring about negative consequences for the missing person, their family, and society in general. Therefore, it is necessary to better understand the phenomenon of missingness and thus find ways to solve cases in the most adequate manner for all parties involved. Age-related cognitive changes and a higher vulnerability to dementia can increase the likelihood of older adults going missing. Thus, the present study sought to identify individual and environmental factors that might predict that an older adult reported missing will be found. To do so, supervised machine learning models were used based on the missing person cases open data of Colombia between 1930 and June, 2021 (*n* = 7855). Classification algorithms were trained to predict whether an older adult who went missing would eventually be found. The classification models with the best performance in the test data were those based on decision trees. Particularly, the Light Gradient Boosting Machine algorithm showed 71% classification accuracy (i.e., 8% above a base model built with the mean of the reported missingness period of the training data). The features with the greatest contribution to the classification were date and place of the missing person case, as well as the age and sex of the missing person. These results help us better understand the societal phenomenon of person missingness and can have important practical implications.
Keywords: *Aging; Classification; Machine Learning; Missing person cases; Older adults*

INTRODUCCIÓN

La desaparición de personas es un fenómeno enigmático y frecuente en muchas sociedades que puede traer consecuencias negativas para la persona desaparecida y

sus familiares, pero también para la sociedad en general. Las causas de la desaparición varían de un caso a otro (ver, *e.g.*, Vargas Rodríguez, 2010) y dependen de las condiciones particulares de la persona (*e.g.*, la edad o estado mental) o las circunstancias específicas de la desaparición (*e.g.*, después de un accidente o catástrofe natural), incluso sin tener en cuenta las desapariciones forzadas, las cuales ocurren ajenas a la voluntad de la persona.

Con la edad ocurren diversos cambios cognitivos en funciones mentales como la atención, la memoria o el control cognitivo (Dobbs & Rule, 1989; McAvinue et al., 2012; Salthouse et al., 1997). Igualmente, con la edad se incrementa el riesgo de depresión y deterioro cognitivo o demencia (Jorm, 2000). Por eso, los adultos mayores pueden ser especialmente vulnerables al fenómeno de la desaparición. Por ejemplo, muchos adultos mayores pueden desaparecer debido a que se perdieron al salir a caminar (Gergerich & Davis, 2017; Neubauer et al., 2021) o, incluso, de manera voluntaria, en el caso de planear y llevar a cabo un suicidio (Vargas Rodríguez, 2010). Además, en caso de una desaparición “no voluntaria” o que involucre un cambio de contexto, una persona mayor puede desorientarse aún más en tiempo o en espacio durante el estado de desaparición. Esta particularidad puede disminuir la probabilidad de encontrar a la persona desaparecida o de que ella pueda volver por sí misma a su contexto original.

Tres dimensiones del comportamiento pueden tipificar la desaparición de una persona adulta: tipo disfuncional, personas con problemas mentales; tipo de escape, personas que deciden o se ven abocadas a desaparecer para ganar independencia o huir de dificultades; y tipo sin intención, aquellas personas que desaparecen bajo la influencia de otra (u otras) o como resultado de un accidente o problema de comunicación con sus allegados (Bonny et al., 2016). Las tipologías que más caracterizan a los adultos mayores, es decir, personas mayores de 60 años, es la disfuncional, seguida por la de escape (Bonny et al., 2016). Esta particularidad, unida a la multiplicidad de circunstancias del entorno asociadas a la situación de desaparición, hacen que las consecuencias de la desaparición puedan ser inocuas (o incluso positivas) pero también fatales y puedan impactar no sólo a la persona desaparecida, sino también a otros directa o indirectamente relacionada con ella (Taylor et al., 2019).

La desaparición de personas causa problemas de salud mental tanto en la persona desaparecida como en sus allegados debido a la incertidumbre de tal situación. En muchos casos, los familiares tienen dificultades para realizar un duelo, aun después de muchos años y de una alta probabilidad de que la persona desaparecida ya haya muerto (Heeke et al., 2015). En el caso de personas mayores, está el agravante de la mayor vulnerabilidad al deterioro de sus capacidades cognitivas y, por consiguiente, una mayor desorientación en espacio, tiempo o, incluso, persona (*e.g.*, en el deterioro cognitivo leve o la demencia). Asimismo, está presente una mayor cantidad de condiciones médicas que, a su vez, hacen más imperioso encontrar a la persona mayor desaparecida. Por eso, es de vital importancia para la sociedad poder entender mejor el fenómeno de la desaparición y así encontrar maneras para que se resuelva de la manera más expedita y favorable para todos.

Numerosos factores pueden hacer más o menos probable que una persona mayor aparezca (Cohen et al., 2008), a partir de la guía que ofrecen tanto a los investigadores encargados de los casos de personas mayores desaparecidas (Fyfe et al., 2015) como a las personas desaparecidas mismas para encontrar su manera de retornar. Por ejemplo, los mayores recursos cognitivos o mentales de una persona o una red social o familiar más amplia podrían aumentar la probabilidad de hallazgo o retorno de la persona mayor en situación de desaparición. Además, los factores del ambiente también pueden tener un papel importante. Por ejemplo, una menor organización del ambiente en donde ocurra la desaparición podría proporcionar menos claves de ayuda para la orientación o búsqueda de la persona mayor, reduciendo así la probabilidad de retorno o hallazgo. De ahí que el objetivo principal del presente trabajo sea pronosticar la probabilidad de que una persona mayor que ha sido reportada como desaparecida aparezca a partir del uso de modelos de aprendizaje automático, o *Machine Learning*, supervisado.

El aprendizaje automático es un método de inteligencia artificial que le permite a una máquina (*i.e.*, un computador) inferir las reglas para construir predicciones de manera automática, sin ser explícitamente programada para tal fin (Géron, 2019; Sen et al., 2020). Las tareas de clasificación del aprendizaje automático se presentan como una herramienta idónea (Chen et al., 2020; Kotsiantis et al., 2006) en el contexto del

estudio de un fenómeno complejo, tal como el de las desapariciones y de la necesidad de la predicción de la posible aparición futura de una persona desaparecida. Algunos trabajos previos han utilizado métodos de aprendizaje automático para estudiar los perfiles de personas desaparecidas o pronosticar la probabilidad de aparición. En este sentido, las primeras investigaciones utilizaron minería de datos para demostrar que se pueden derivar reglas que permiten predecir el resultado de casos de personas desaparecidas y, así, sustentar las intuiciones de los investigadores de la policía (Blackmore et al., 2005). Incluso, propuestas futuras apuntan también hacia la utilización de modelos de aprendizaje automático para la búsqueda de personas desaparecidas (Pedroza Manga, 2019).

Un estudio reciente con una muestra de personas de todas las edades desaparecidas en Colombia durante 2017 mostró un adecuado desempeño de modelos de aprendizaje *K-Nearest Neighbours* y árboles de decisión para pronosticar si una persona aparece viva *vs.* muerta y aparece *vs.* no aparece, respectivamente (Delahoz-Domínguez & Mendoza-Brand, 2021). Otro estudio en una muestra similar, que buscó identificar perfiles de personas desaparecidas utilizando *Waikato Environment for Knowledge Analysis* (o WEKA), encontró algunos perfiles que sugieren que en Bogotá, los adolescentes, las mujeres solteras y los usuarios de sustancias psicoactivas parecen desaparecer de manera voluntaria, mientras que las desapariciones forzadas parecen ocurrir principalmente en la ciudad de Ibagué (Rolong Agudelo et al., 2020). Sin embargo, a pesar de las características particulares de las personas mayores desaparecidas y de su vulnerabilidad, hasta la fecha no se ha abordado el estudio de la adultez mayor en situación de desaparición o por causas diferentes a la desaparición forzada en Colombia en los últimos 50 años.

En suma, en el presente estudio se busca identificar los factores individuales y del entorno que pueden pronosticar la aparición de una persona mayor con la ayuda de modelos de aprendizaje automático supervisado. Para tal fin, se realizará el análisis de los datos abiertos proporcionados por el Sistema de Información Red de Desaparecidos y Cadáveres (SIRDEC) del Instituto Nacional de Medicina Legal y Ciencias Forenses a través de Datos Abiertos Colombia. Específicamente, se buscará encontrar la probabilidad de que una persona mayor aparezca a través de modelos

supervisados de clasificación, así como identificar qué características de la persona o de su entorno aportan a dicha probabilidad a través de modelos interpretativos de aprendizaje automático.

METODOLOGÍA

Datos

El presente estudio se basó en los datos abiertos proporcionados por el SIRDEC del Instituto Nacional de Medicina Legal y Ciencias Forenses de Colombia a través de Datos Abiertos Colombia disponibles en el sitio web: (https://www.datos.gov.co/Justicia-y-Derecho/Desaparecidos-Colombia-hist-rico-a-os-1930-a-junio/8hqm-7fdt), los cuales fueron descargardos el 5 de agosto de 2021. La versión original de la base de datos contaba con 162.401 registros de personas desaparecidas desde el año 1930 hasta el año 2021. Las fases para llevar a cabo el presente estudio fueron: (i) limpieza de la base de datos y selección de los registros relevantes, (ii) realización de análisis descriptivos y (iii) identificación de posibles modelos y la respectiva evaluación de su desempeño.

Preparación inicial de los datos y las variables

En la primera fase, se eliminaron los registros sin información en las variables de edad (n = 202) y fecha de desaparición (n = 129), ya que se consideraba fundamental determinar que el registro correspondiera al de una persona mayor al momento de la desaparición y porque la fecha de la desaparición puede ser uno de los factores más relevantes en el pronóstico de aparición. Adicionalmente, se eliminaron registros cuya causa de desaparición fuera presuntamente forzada (n = 32.403), en concordancia con el planteamiento del problema, ya que esta causa dificulta encontrar patrones de pronóstico de aparición al depender de factores externos a la persona desaparecida misma y ser de carácter más complejo. Luego de esta limpieza general y con base en los objetivos del presente estudio, se excluyeron los registros cuya edad al momento de la desaparición fuera menor a 60 años, cuyo status de desaparición fuera “apareció muerto” y cuyo país de desaparición no fuera Colombia, lo cual dejó n = 7855 registros válidos. Las variables predictoras que se tuvieron en cuenta fueron la fecha y el

municipio de la desaparición – como variables externas a la persona – y la edad, el sexo, el estado civil, la escolaridad y el factor de vulnerabilidad – como variables inherentes a la persona desaparecida. Las variables "país de nacimiento" y "ancestro racial" no se tuvieron en cuenta porque no se hipotetizan como relevantes para la predicción y porque no presentaron mayor variabilidad en su distribución. En la última parte de esta fase se realizó la transformación de variables con el fin de su adecuación para el entrenamiento del modelo de aprendizaje automático (Tabla 1). Posteriormente, en la segunda fase, se realizaron los respectivos análisis descriptivos de cada variable, con el fin de identificar la distribución de los datos, así como los valores faltantes.

Tabla 1. Transformación de algunas de las variables previa al análisis

Variable original	**Nueva variable**	**Categorías de la nueva variable = categorías de la variable original**
Objetivo		
Status	Apareció	0 = "Desaparecido" 1 = "Apareció Vivo"
Predictoras		
Fecha	Duración (días)	Número de días hasta el 30 de julio de 2021 = Fecha de la desaparición
Estado civil	Relación conyugal	Actual = "Unión libre", "Casado (a) Pasada = "Separado (a)", "Divorciado (a)", "Viudo (a)" Ninguna = "Soltero (a)"
Escolaridad	Años de escolaridad	0.0 = "Sin escolaridad" 2.5 = "Educación inicial y educación preescolar" 5.0 = "Educación básica primaria" 7.5 = "Educación básica secundaria o secundaria baja" 10.0 = "Educación media o secundaria alta" 12.5 = "Educación técnica profesional y tecnológica" 15.0 = "Universitario" 17.5 = "Especialización, Maestría o equivalente" 20.0 = "Doctorado o equivalente"
Factor de vulnerabilidad	Vulnerabilidad	No = "Ninguno" Sí = Todos los valores excepto "Sin información"
Municipio y departamento de desaparición	Municipio (habitantes)*	Número de habitantes (entre 2015 y 2018) a partir de la información en la página de Wikipedia (https://es.wikipedia.org/wiki/Municipios_de_Colombia) y sus respectivos anexos (*e.g.*, https://es.wikipedia.org/wiki/Anexo:Municipios_de_Huila)

* Se calculó con ayuda de *Web-scraping* [https://github.com/virtualmarioe/Web_scraping_tutorial]

Preprocesamiento y modelación

En la tercera fase del análisis, se realizó la preparación requerida para la modelación así como la modelación misma. Específicamente, en cuanto a la preparación, primero se separaron los datos en datos de entrenamiento y prueba a razón de 80% (n = 6284) y 20% (n = 1571), respectivamente; después se imputaron los datos faltantes tanto en los sets de entrenamiento como de prueba, con la media de los datos de entrenamiento para las variables numéricas con datos faltantes (*i.e.*, escolaridad y municipio) y con la moda de los datos de entrenamiento para las variables categóricas que tuvieran datos faltantes (*i.e.*, vulnerabilidad y relación conyugal). Dicha imputación se realizó con la función *SimpleImputer*, la cual se ajustó en los datos de entrenamiento solamente y después se aplicó a ambos conjuntos de datos de entrenamiento y de prueba.

Posteriormente, se creó un modelo base y se realizó el ajuste de variables numéricas y categóricas con las funciones *Standard Scaler* y *One Hot Encoder*, respectivamente, y de la variable objetivo con la función *Label Encoder*. Dicho ajuste de variables se realizó solamente en los datos de entrenamiento y después se aplicó tanto a los datos de entrenamiento como a los datos de prueba. De la base de datos completa, 66% (n = 5166) de los registros correspondió a adultos mayores con estatus de “desaparecido” mientras que 34% (n = 2689) correspondió a adultos mayores con status de “aparecido”. Por eso, se utilizaron dos técnicas para corregir el desbalance en la proporción entre las clases durante el entrenamiento del modelo: (a) la técnica SMOTE (*Synthetic Minority Oversampling Technique*), la cual permitió aumentar la clase minoritaria a partir del aumento sintético de los datos (n_1 = n_2 = 4133) y (b) la técnica de sub-muestreo de los datos de entrenamiento para conseguir un balance de clases 50 / 50 (n_1 = n_2 = 2151).

En cuanto a la modelación, primero, se realizó un análisis global de posibles modelos de clasificación para la predicción de la categoría de salida “aparece” (0 = “no”, 1 = “sí”), de los cuales se seleccionaron los tres modelos con los puntajes de *Accuracy* o exactitud (*i.e.*, número de predicciones correctas / número total de predicciones) más altos. Posteriormente, se examinaron las matrices de confusión de

los tres mejores modelos, así como diversas métricas de su desempeño, tales como: *Recall* o sensibilidad (*i.e.*, identificación de casos positivos verdaderos de todos los posibles casos positivos), *Precision* o especificidad (*i.e.*, identificación de casos positivos verdaderos de todos los casos identificados como positivos) y el puntaje *F1* (*i.e.*, media armónica ponderada entre la sensibilidad y la especificidad). Las tres fases del análisis se realizaron en Python (v. 3.6) con los paquetes PyCaret (https://pycaret.org/) y Scikit Learn (https://scikit-learn.org/stable/) y se pueden consultar en el siguiente enlace [https://github.com/alruizzo/missing_persons].

RESULTADOS

La edad promedio fue de 71.35 ± 8.36 años en los casos de personas que aparecieron vs. 71.45 ± 9.91 en personas que continuaron desaparecidas (Figura 1) y la escolaridad promedio fue de 5.12 ± 3.53 vs. 4.85 ± 3.39 años, respectivamente. La mayoría de casos fueron de sexo masculino (72.8% de los registros con estatus de aparecido vs. 83% de los registros con estatus de desaparecido) y correspondió a personas sin ningún factor de vulnerabilidad evidente (74.3% estatus aparecido vs. 71.7% estatus desaparecido) y con una relación actual (40.2% estatus aparecido vs. 49.2% estatus desaparecido) al momento de la desaparición. Casi la mitad de las desapariciones ocurrieron en municipios de menos de 1 millón de habitantes y casi un 36% ocurrieron en la capital (con aprox. 8 millones de habitantes), con una proporción mayor de personas con estatus aparecido en municipios con población superior a los 2 millones de habitantes (Figura 2). La mayoría de los casos se reportaron hace menos de 5000 días (aprox. 14 años), siendo este tiempo el límite superior para casi el 100% de los casos de personas con estatus aparecido (Figura 3).

Siguiendo lo observado en el análisis descriptivo, el modelo base consistió en utilizar la media del número de días desde el reporte de desaparición como regla de predicción de la aparición. Este valor, de 4474.8 días, se calculó en los datos de entrenamiento y se utilizó en los datos de prueba, con lo cual se consiguió un 63% de *Accuracy* (Tabla 2). Este valor se utilizó para comparar el desempeño relativo de los modelos de aprendizaje automático y poder así juzgar su utilidad. Como se puede apreciar en la Tabla 2, el desempeño de los “mejores” modelos – todos basados en

modelos de árboles de decisión – fue similar entre sí en todas las métricas. Sin embargo, se seleccionó el modelo de máquina de refuerzo de gradiente ligero (*Light Gradient Boosting Machine,* LGBM, por sus siglas en inglés), ya que tuvo los valores más altos en la exactitud de la clasificación o *Accuracy* (los valores detallados de todos los modelos se pueden encontrar en Anexos: Figura S1).

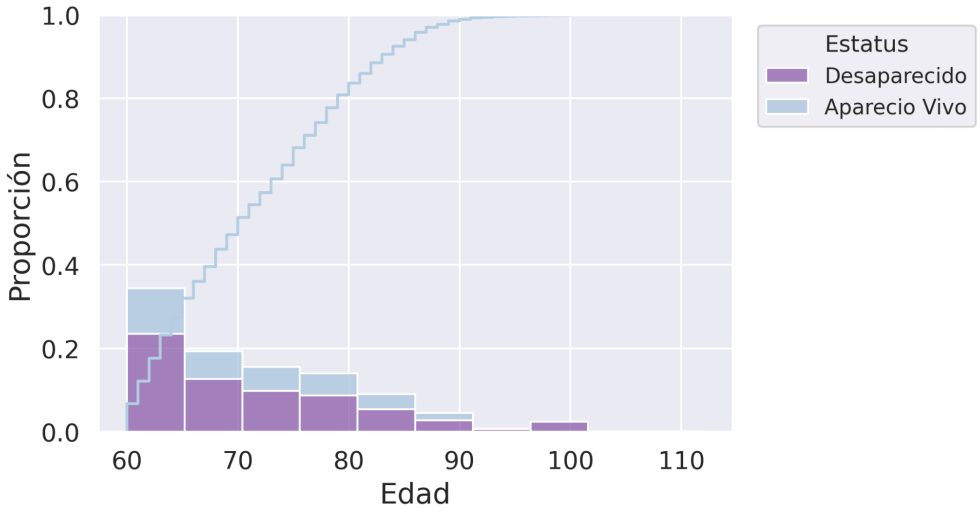

Figura 1. Histograma de edad (en años) por estatus de desaparición de la base de datos completa (*n* = 7855). La distribución de edad fue similar en ambos grupos. La línea azul sobre las barras del histograma representa la función de distribución empírica acumulada o la proporción de observaciones que están debajo de cada valor único en el conjunto de datos.

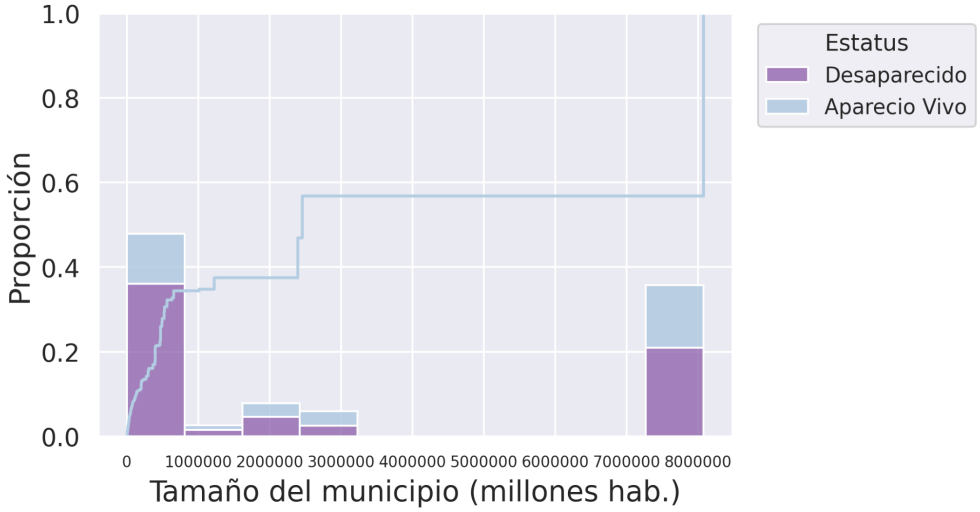

Figura 2. Histograma del tamaño del municipio (en millones de habitantes) donde ocurrió la desaparición por estatus de desaparición (*i.e.*, “desaparecido” o “apareció vivo”). La línea azul sobre las barras indica la proporción acumulada de los registros con estatus aparecido: la mayor proporción de casos con estatus de aparecido se encuentra en municipios por encima de los dos millones de habitantes.

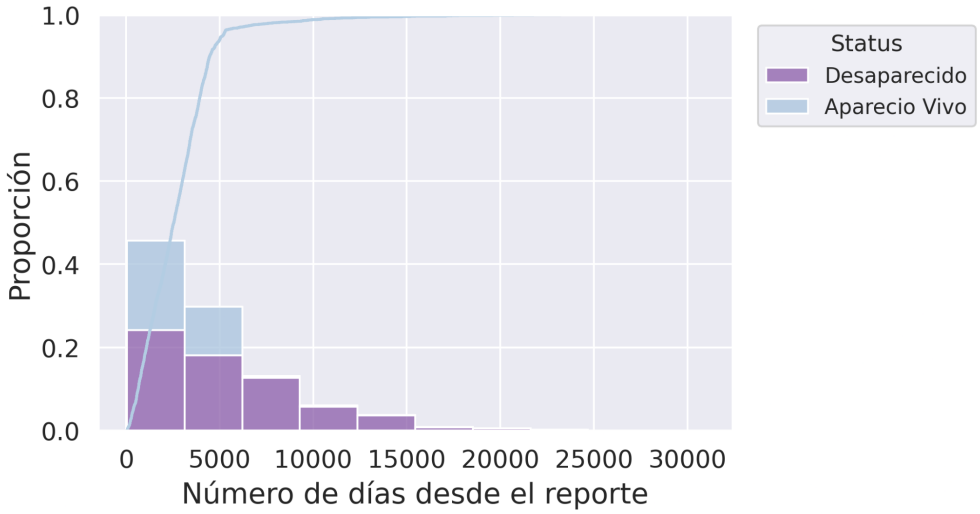

Figura 3. Histograma de la fecha de la desaparición (en número de días desde el reporte hasta el 30 de julio de 2021) por estatus de desaparición (*i.e.*, "desaparecido" o "apareció vivo"). La línea azul sobre las barras indica la proporción acumulada de los registros con estatus aparecido: más del 90% de los casos de personas con estatus apareció vivo tiene una fecha de registro inferior a 5000 días (14 años aproximadamente).

Tabla 2. Desempeño promedio (con validación cruzada en 10 partes o *folds*) de los mejores modelos en los datos de prueba

Modelo	Accuracy	AUC	Recall	Precisión	F1
Light Gradient Boosting Machine	0.71	0.78	0.67	0.56	0.61
Sin ajuste del balance de clases*	0.70	0.66	0.51	0.57	0.54
Con sub-muestreo*	0.68	0.70	0.78	0.52	0.63
Random Forest Classifier	0.70	0.76	0.61	0.56	0.58
Gradient Boosting Classifier	0.69	0.79	0.78	0.53	0.63
Extra Trees Classifier	0.68	0.73	0.57	0.53	0.55
Ada Boost Classifier	0.68	0.77	0.82	0.52	0.63
Base[a]	0.63	0.69	0.89	0.48	0.62

[a] Basado en la regla de la media de días desde el reporte (4474.8 días) * Sin validación cruzada

El modelo LGBM se exploró en más detalle utilizando un sub-muestreo de los datos de entrenamiento, ya que, con la técnica SMOTE, el modelo penalizaba la clase inicialmente minoritaria (*i.e.*, "aparecido"; Anexos: Figura S2). Como puede verse en la

matriz de confusión (Figura 4), 22% de los registros fueron falsos negativos (*i.e.*, se predijo que continuaban desaparecidos pero aparecieron), mientras que 37% de los registros fueron falsos positivos (*i.e.*, se predijo que aparecían pero continuaron desaparecidos), lo cual representó una mejoría sustancial con respecto al modelo base, en el cual estos porcentajes fueron 11% y 50%, respectivamente (Anexos: Figura S3). Adicionalmente, con respecto al modelo utilizando SMOTE, el modelo con datos de entrenamiento sub-muestreados mostró un porcentaje más alto de sensibilidad (*Recall*) en la clase "aparecido" (0.78 vs. 0.65 en el de SMOTE) y de F1 (0.63 vs. 0.60 en el de SMOTE) (Anexos: Figura S4).

Uno de los objetivos principales del estudio fue la identificación de los factores que determinan la aparición de una persona mayor reportada como desaparecida en Colombia. En consonancia con este objetivo, se revisaron las categorías más importantes que contribuyeron a la predicción en el modelo LGBM (Figura 5 y Anexos: Figura S5). Éstas fueron: número de días desde el reporte de la desaparición, tamaño del municipio (en número de habitantes) donde ocurrió la desaparición, edad al momento de la desaparición, y, ligeramente, sexo. Algunos ejemplos de los valores de estas variables y de las predicciones específicas pueden verse en los Anexos (Figura S6).

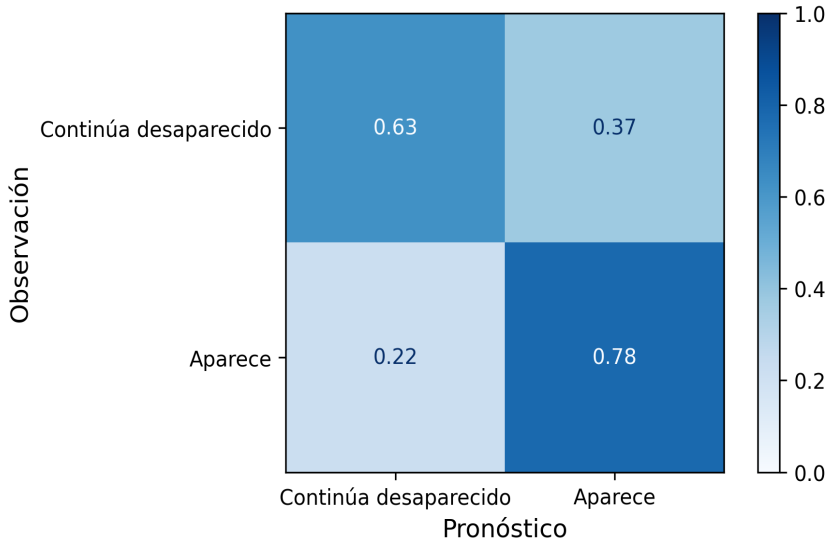

Figura 4. Matriz de confusión modelo LGBM con datos de entrenamiento sub-muestreados. Los porcentajes de clasificación se muestran normalizados por fila. Para la clase "Aparece", 78% se

pronostican correctamente, mientras que para la clase “Desaparecido”, 63% se pronostican correctamente.

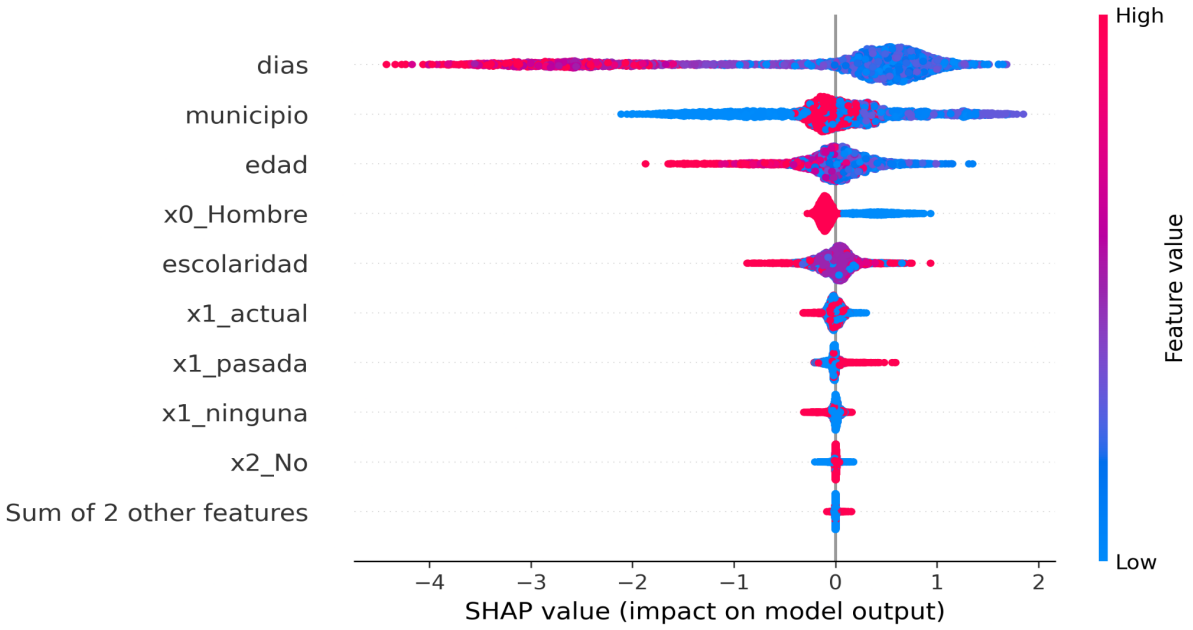

Figura 5. Importancia de las categorías utilizadas para la predicción. Las categorías utilizadas para la predicción se muestran en orden de importancia en el eje vertical, mientras que los valores SHAP (*SHapley Additive exPlanations*) se muestran en el eje horizontal, con valores negativos representando la etiqueta “desaparecido” y los valores positivos la etiqueta “aparecido”. Cada punto es un registro de los datos de prueba. La escala de colores codifica el valor del registro, así: puntos azules, valores bajos; puntos morados, valores intermedios; puntos rojos, valores altos. “x0”, “x1” y “x2” = *One Hot Encoding* para las variables “sexo”, “relación conyugal” y “vulnerabilidad”, respectivamente. *Feature value*: valor de la variable; *High*: valores altos; *Low*: valores bajos.

DISCUSIÓN

El presente estudio buscó identificar los factores individuales y del entorno que pueden pronosticar la aparición de una persona mayor con la ayuda de modelos de aprendizaje automático supervisado. Los resultados mostraron que los mejores modelos fueron aquellos basados en árboles de decisión; en especial, el *light gradient boosting machine*, el *random forest* y el *gradient boosting classifier*. Los niveles de error de estos modelos (29%, 30% y 31%, respectivamente) estuvieron por debajo del nivel de error de un modelo base que utilizó como regla de predicción la media de los días transcurridos desde el reporte de la desaparición en los datos de entrenamiento (*i.e.*, 37%). Este hallazgo indica que los modelos de aprendizaje automático pueden brindar

una mayor información sobre los factores del pronóstico de aparición de personas mayores reportadas como desaparecidas así como sobre el pronóstico mismo. Los factores cruciales identificados como más relevantes para la aparición fueron un menor número de días desde la desaparición, una menor edad de la persona desaparecida, una ciudad donde ocurre la desaparición con mayor número de habitantes y, ligeramente, sexo femenino. Los presentes resultados pueden ayudarnos a entender mejor el fenómeno de la desaparición en personas mayores al tiempo que pueden tener implicaciones prácticas (Anexos: Figura S7).

Los modelos más exactos en la clasificación en el presente estudio fueron modelos basados en árboles de decisión, en consonancia con algunas investigaciones previas (*e.g.*, Blackmore et al., 2005; Delahoz-Domínguez & Mendoza-Brand, 2021). Específicamente, en el presente trabajo se seleccionó como mejor modelo al LGBM, el cual es una implementación especial del algoritmo *Gradient Boosting Decision Tree* (Ke et al., 2017). Este modelo, entrenado con datos balanceados a partir de un sub-muestreo de la clase dominante ("desaparecido"), nos permitió maximizar la medida de *recall* o sensibilidad en ambas clases con respecto al modelo base. Este resultado implica que nuestro modelo logró reducir la tasa de falsos positivos (*i.e.*, predecir que un caso aparece cuando en realidad continúa desaparecido) de un 50% a un 37% con respecto al modelo base.

Nuestro modelo también logró identificar los aspectos clave en la predicción de la aparición. Como se había esperado, factores intrínsecos a la persona como su edad , la cual se relaciona con el estado cognitivo (Dobbs & Rule, 1989; Jorm, 2000; McAvinue et al., 2012; Salthouse et al., 1997; Whalley et al., 2004), o el sexo, el cual se puede relacionar con el motivo de la desaparición (García-Barceló et al., 2020) o el tipo de comportamientos durante la misma, fueron relevantes. Por otro lado, factores extrínsecos como la fecha de la desaparición o el tamaño del municipio donde ocurrió la desaparición, los cuales se relacionan de manera indirecta con la estructura y organización del entorno físico y social, también demostraron ser cruciales para la predicción.

Contrario a lo esperado, otros factores intrínsecos como la presencia de vulnerabilidad, la relación conyugal o el nivel de escolaridad de la persona

desaparecida, no contribuyeron de manera significativa a la predicción. Una explicación para estos hallazgos negativos es la relativa baja variabilidad de los datos (sumada a la alta proporción de datos faltantes) en estas categorías. Por eso, en el futuro, la cuantificación de estas variables (*e.g.*, número de personas con las que convive la persona desaparecida o el número de características de vulnerabilidad de la persona) o un mayor detalle en general (*e.g.*, los años efectivos de estudio o la persona que realiza el reporte) podría ayudar a dilucidar si estás categorías tienen un impacto sobre la probabilidad de aparición.

Los resultados del presente trabajo deben mirarse a la luz de algunas limitaciones. Por ejemplo, los datos sobre los cuales se realizaron los análisis, debido a su naturaleza, no son datos recolectados con fines de investigación científica y, por lo tanto, no incluyen toda la información en la profundidad o el detalle que se requeriría si estuviera guiada por la teoría. Adicionalmente, la cantidad de valores faltantes es notable y tuvo que resolverse a partir de métodos de imputación, los cuales añaden cierto grado de incertidumbre a las predicciones. Finalmente, aún queda por determinar si los presentes resultados y conclusiones se generalizan a casos de personas menores de 60 años, donde la desaparición es forzosa o donde el desenlace es fatal. De todas maneras, a pesar de sus limitaciones, este estudio sienta un precedente para el mejor entendimiento del fenómeno de la desaparición en adultos mayores en Colombia.

CONCLUSIÓN

En el presente estudio se lograron identificar los factores individuales (como la edad y el sexo) y del entorno (como el tiempo y espacio que enmarcan la desaparición) que pronostican la aparición de una persona mayor, con la ayuda de un modelo de aprendizaje automático supervisado basado en árboles de decisión. Tal identificación de factores no hubiera sido posible con un modelo basado en una media aritmética. Adicionalmente, el modelo propuesto en el presente trabajo logra reducir el error del modelo base en un 5% y reducir la tasa de falsos positivos en un 13%, lo cual puede tener implicaciones prácticas.

ANEXOS

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7086	0.7818	0.6695	0.5630	0.6114	0.3812	0.3851	0.225
rf	Random Forest Classifier	0.6999	0.7645	0.6128	0.5560	0.5828	0.3493	0.3505	1.164
gbc	Gradient Boosting Classifier	0.6914	0.7863	0.7834	0.5339	0.6348	0.3840	0.4055	0.686
et	Extra Trees Classifier	0.6824	0.7335	0.5658	0.5346	0.5494	0.3046	0.3051	0.993
ada	Ada Boost Classifier	0.6773	0.7692	0.8206	0.5181	0.6351	0.3713	0.4032	0.331
dummy	Dummy Classifier	0.6577	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.057
dt	Decision Tree Classifier	0.6558	0.6305	0.5504	0.4979	0.5225	0.2546	0.2556	0.083
knn	K Neighbors Classifier	0.6431	0.6983	0.6564	0.4842	0.5572	0.2695	0.2784	0.179
ridge	Ridge Classifier	0.6364	0.0000	0.8001	0.4813	0.6010	0.3031	0.3367	0.062
lda	Linear Discriminant Analysis	0.6364	0.7075	0.8001	0.4813	0.6010	0.3031	0.3367	0.071
lr	Logistic Regression	0.6361	0.7059	0.7620	0.4802	0.5890	0.2915	0.3168	0.430
qda	Quadratic Discriminant Analysis	0.6360	0.6870	0.0898	0.1619	0.0540	0.0077	0.0163	0.065
nb	Naive Bayes	0.6252	0.7079	0.7940	0.4719	0.5919	0.2848	0.3183	0.062
svm	SVM - Linear Kernel	0.6173	0.0000	0.8675	0.4682	0.6073	0.2940	0.3522	0.086

Figura S1. Modelos de clasificación explorados. Tabla de resumen con todos los modelos de clasificación explorados usando PyCaret.

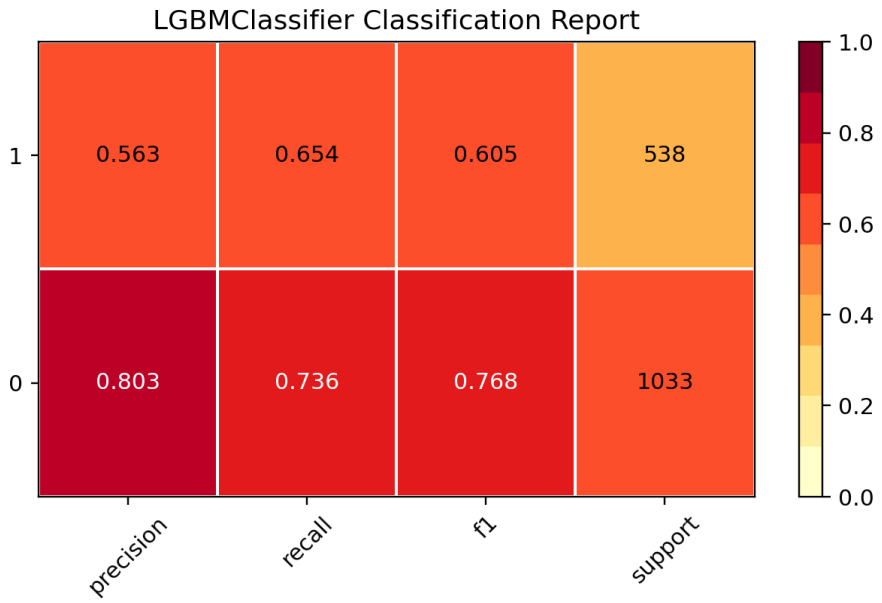

Figura S2. Reporte de clasificación modelo *Light Gradient Boosting Machine* (LGBM). Reporte de clasificación separado por clases (0 = "Desaparecido", 1 = "Aparece") en los datos de prueba (n_0 = 1033,

n_1 = 538). La escala de colores representa la escala de los resultados. Columnas, respectivamente de izquierda a derecha: precisión, sensibilidad, puntaje F1 y número de muestras por clase.

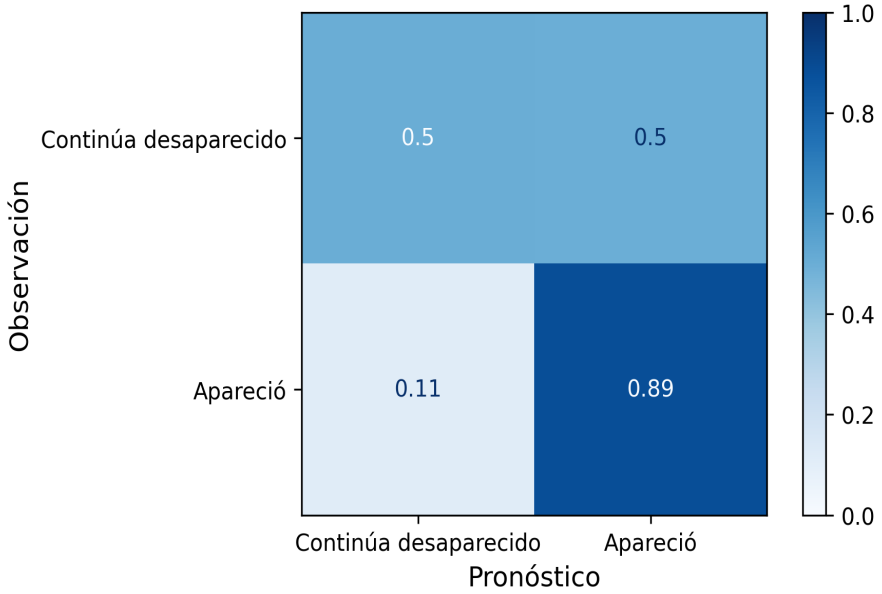

Figura S3. Matriz de confusión del modelo base. La predicción fue excelente para la clase "apareció" (89%), pero estuvo al nivel de chance para la clase de "desaparecido". Los porcentajes son con referencia a las filas de la matriz.

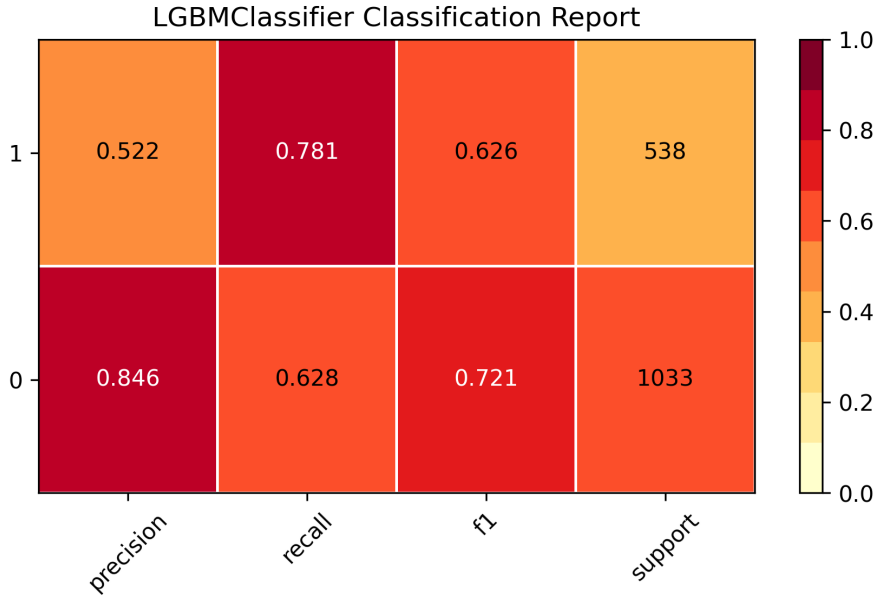

Figura S4. Reporte de clasificación modelo LGBM con datos de entrenamiento sub-muestreados. Véase la Figura S1 para comparación con la técnica SMOTE (*oversampling*) en el muestreo.

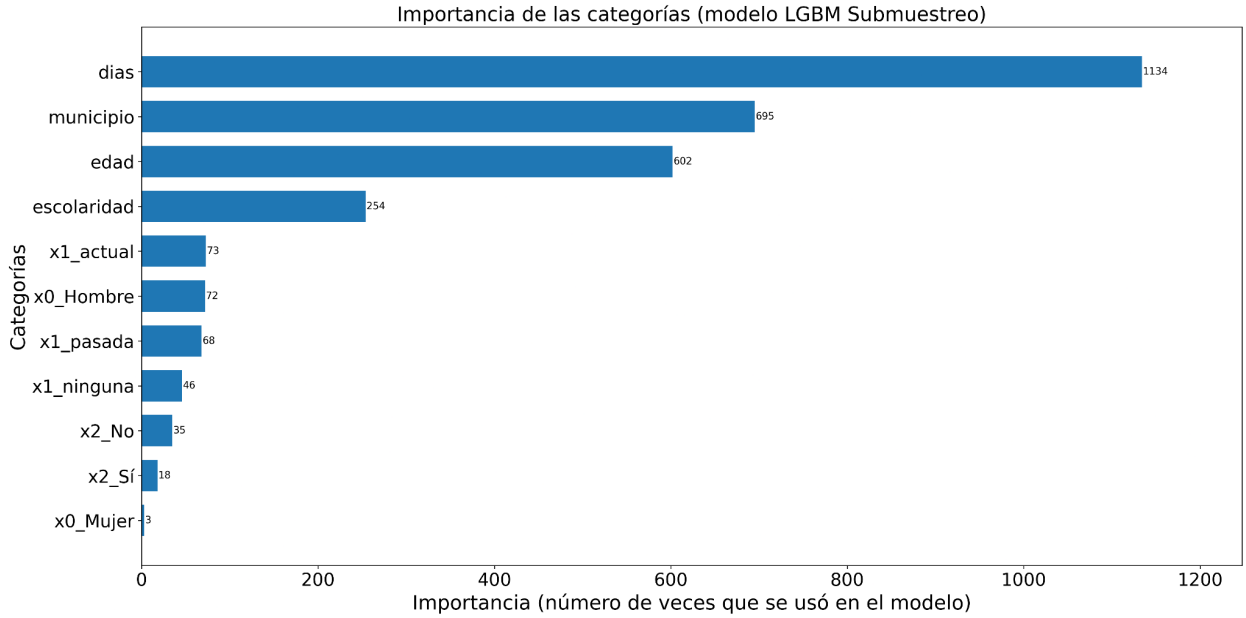

Figura S5. Gráfico de barras sobre la importancia de las categorías. Las categorías utilizadas para el pronóstico del modelo LGBM se muestran en orden de importancia (eje vertical), según el número de veces que se utilizaron para la predicción (eje horizontal).

edad	sexo	escolaridad	vulnerabilidad	municipio	dias	relacion	prob_desaparec_pred	prob_aparec_pred	aparecio
63	Mujer	4.966056	Sí	36708.0	7872	ninguna	0.944344	0.055656	0
100	Hombre	4.966056	No	491387.0	6451	ninguna	0.977852	0.022148	0
71	Hombre	5.000000	No	18678.0	7724	actual	0.937428	0.062572	0
72	Hombre	4.966056	No	2394870.0	1576	ninguna	0.307530	0.692470	0
60	Hombre	15.000000	No	2457680.0	6672	actual	0.861568	0.138432	0
...	...	...	...	...	...	...	...	...	...
63	Mujer	4.966056	No	495200.0	2941	ninguna	0.418575	0.581425	1
77	Hombre	10.000000	Sí	8076734.0	2339	actual	0.487001	0.512999	0
86	Mujer	0.000000	No	8076734.0	2882	pasada	0.195247	0.804753	1
62	Hombre	4.966056	No	562704.0	9467	actual	0.749354	0.250646	1
80	Hombre	4.966056	No	11147.0	4398	pasada	0.942365	0.057635	0

Figura S6. Captura de pantalla de probabilidades pronosticadas para algunos ejemplos de los datos de prueba. Esta figura se presenta para facilitar el entendimiento del impacto de las categorías en los pronósticos de los datos de prueba. Las primeras 7 columnas son las variables de predicción con los valores específicos para cada fila; las últimas tres columnas representan, en orden: la probabilidad pronosticada de que ese caso continúe desaparecido, la probabilidad pronosticada de que ese caso aparezca y, por último, el valor real de si ese caso continúa desaparecido (“0”) o apareció (“1”).

edad	sexo	escolaridad	vulnerabilidad	municipio	dias	relacion	prob_desaparec_pred	prob_aparec_pred	aparecio
69	Hombre	2.500000	No	5.348000e+03	152	actual	0.395853	0.604147	1
73	Hombre	10.000000	No	8.076734e+06	150	actual	0.295671	0.704329	0
91	Hombre	5.000000	No	8.076734e+06	150	actual	0.594669	0.405331	1
76	Hombre	7.500000	No	8.076734e+06	149	ninguna	0.428235	0.571765	0
61	Hombre	5.000000	No	8.076734e+06	152	actual	0.389177	0.610823	1
76	Hombre	5.000000	No	8.076734e+06	152	pasada	0.524417	0.475583	0
61	Hombre	10.000000	No	8.076734e+06	148	ninguna	0.503226	0.496774	0
73	Mujer	5.000000	No	8.076734e+06	147	pasada	0.333359	0.666641	1
72	Hombre	0.000000	No	8.076734e+06	147	actual	0.389371	0.610629	0
74	Hombre	4.966056	No	1.220300e+05	152	actual	0.639750	0.360250	0
66	Mujer	5.000000	No	8.076734e+06	145	ninguna	0.273307	0.726693	0
74	Hombre	10.000000	No	4.720230e+05	144	actual	0.149899	0.850101	1
65	Hombre	12.500000	No	1.223967e+06	148	actual	0.310915	0.689085	0
64	Hombre	4.966056	No	3.359334e+06	143	actual	0.347027	0.652973	1
62	Hombre	10.000000	No	8.076734e+06	142	pasada	0.243117	0.756883	0
65	Hombre	12.500000	No	1.698400e+04	143	ninguna	0.509335	0.490665	1
76	Hombre	4.966056	No	3.974880e+05	152	pasada	0.231798	0.768202	1
61	Hombre	12.500000	No	8.076734e+06	140	actual	0.588837	0.411163	0

Figura S7. Captura de pantalla de probabilidades pronosticadas para algunos ejemplos de datos nunca antes vistos. Nuevos casos ocurridos entre julio y septiembre de 2021 con los cuales se generaron probabilidades de aparición (penúltima columna, “prob_aparec_pred”) y valores reales de aparición hasta noviembre de 2021 (0 = desaparecido; 1 = aparece).

REFERENCIAS BIBLIOGRÁFICAS

Blackmore, K., Bossomaier, T., Foy, S., & Thomson, D. (2005). Data Mining of Missing Persons Data. In S. K. Halgamuge & L. Wang (Eds.), *Classification and Clustering for Knowledge Discovery* (pp. 305–314). Springer. https://doi.org/10.1007/11011620_19

Bonny, E., Almond, L., & Woolnough, P. (2016). Adult Missing Persons: Can an Investigative Framework be Generated Using Behavioural Themes? *Journal of Investigative Psychology and Offender Profiling*, *13*(3), 296–312. https://doi.org/10.1002/jip.1459

Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Cohen, I. M., McCormick, A. V., & Plecas, D. (2008). *A Review of the Nature and Extent of*

Uncleared Missing Persons Cases in British Columbia [University College of the Fraser Valley]. https://ufv.ca/media/assets/ccjr/reports-and-publications/Missing_Persons.pdf

Delahoz-Domínguez, E., & Mendoza-Brand, S. (2021). A predictive model for the missing people problem. *Romanian Journal of Legal Medicine*, *29*(1), 74–80. https://doi.org/10.4323/rjlm.2021.74

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500–503. https://doi.org/10.1037/0882-7974.4.4.500

Fyfe, N. R., Stevenson, O., & Woolnough, P. (2015). Missing persons: The processes and challenges of police investigation. *Policing and Society*, *25*(4), 409–425. https://doi.org/10.1080/10439463.2014.881812

García-Barceló, N., González Álvarez, J. L., Woolnough, P., & Almond, L. (2020). Behavioural themes in Spanish missing persons cases: An empirical typology. *Journal of Investigative Psychology and Offender Profiling*, *17*(3), 349–364. https://doi.org/10.1002/jip.1562

Gergerich, E., & Davis, L. (2017). Silver Alerts: A Notification System for Communities with Missing Adults. *Journal of Gerontological Social Work*, *60*(3), 232–244. https://doi.org/10.1080/01634372.2017.1293757

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Heeke, C., Stammel, N., & Knaevelsrud, C. (2015). When hope and grief intersect: Rates and risks of prolonged grief disorder among bereaved individuals and relatives of disappeared persons in Colombia. *Journal of Affective Disorders*, *173*, 59–64. https://doi.org/10.1016/j.jad.2014.10.038

Jorm, A. F. (2000). Is depression a risk factor for dementia or cognitive decline? A review. *Gerontology*, *46*(4), 219–227. https://doi.org/10.1159/000022163

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM:

A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

McAvinue, L. P., Habekost, T., Johnson, K. A., Kyllingsbæk, S., Vangkilde, S., Bundesen, C., & Robertson, I. H. (2012). Sustained attention, attentional selectivity, and attentional capacity across the lifespan. *Attention, Perception, & Psychophysics*, *74*(8), 1570–1582. https://doi.org/10.3758/s13414-012-0352-6

Neubauer, N., Daum, C., Miguel-Cruz, A., & Liu, L. (2021). Mobile alert app to engage community volunteers to help locate missing persons with dementia. *PLOS ONE*, *16*(7), e0254952. https://doi.org/10.1371/journal.pone.0254952

Pedroza Manga, R. E. (2019). *Diseño e implementación de un sistema de biometría facial para la búsqueda e identificación de personas desaparecidas en Colombia*. Universidad de Cartagena.

Rolong Agudelo, G. E., Montenegro Marin, C., & Gaona García, P. A. (2020). Aplicación de la minería de datos para la detección de perfiles de personas desaparecidas en Colombia. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *E35*, 84–95.

Salthouse, T. A., Toth, J. P., Hancock, H. E., & Woodard, J. L. (1997). Controlled and Automatic Forms of Memory and Attention: Process Purity and the Uniqueness of Age-Related Influences. *The Journals of Gerontology: Series B*, *52B*(5), P216–P228. https://doi.org/10.1093/geronb/52B.5.P216

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (pp. 99–111). Springer. https://doi.org/10.1007/978-981-13-7403-6_11

Taylor, C., Woolnough, P. S., & Dickens, G. L. (2019). Adult missing persons: A concept analysis. *Psychology, Crime & Law*, *25*(4), 396–419. https://doi.org/10.1080/1068316X.2018.1529230

Vargas Rodríguez, P. (2010). Tras las huellas de los desaparecidos “voluntarios” en Bogotá [BachelorThesis, Universidad del Rosario]. In *Instname:Universidad del Rosario*. https://repository.urosario.edu.co/handle/10336/1778

Whalley, L. J., Deary, I. J., Appleton, C. L., & Starr, J. M. (2004). Cognitive reserve and the neurobiology of cognitive aging. *Ageing Research Reviews*, *3*(4), 369–382. https://doi.org/10.1016/j.arr.2004.05.001