Análisis Estadístico Para La Predicción De La Probabilidad De Encontrar Adultos Mayores Desaparecidos Mediante La Aplicación De Modelos De Aprendizaje Automático

Departamento de Ingeniería y Ciencias Básicas, Fundación Universitaria Los Libertadores, Bogotá, Colombia

Información del autor: José John Fredy González Veloza

Codigo: 202015002609

Director: Manuel Francisco Romero Ospina

Abreviaturas: GBC = clasificador potenciador de gradiente; LGBM = máquina potenciadora de gradiente ligero; SIRDEC = Sistema de Información Red de Desaparecidos y Cadáveres (sistema de información de la red de personas desaparecidas y cadáveres); SMOTE = técnica de sobremuestreo minoritario sintético

Agradecimientos

Los autores agradecen a Datos Abiertos Colombia y al Sistema de Información Red de Desaparecidos y Cadáveres por permitirnos el acceso a los datos.

Contribuciones de los autores

José John Fredy González Veloza: Metodología, Software, Supervisión, Redacción – Revisión y edición. **Adriana L. Ruiz-Rizzo:** Conceptualización, Curación de datos, Análisis formal, Adquisición de fondos, Metodología, Software, Visualización, Escritura - borrador original, Escritura - Revisión y edición; **Mario E. Archila-Meléndez**: Recursos, Redacción – Revisión y edición;

Resumen

La desaparición de personas es un fenómeno enigmático y frecuente que puede traer consecuencias negativas para la persona desaparecida, su familia y la sociedad en general. Los cambios cognitivos relacionados con la edad y una mayor vulnerabilidad a la demencia pueden aumentar la propensión de los adultos mayores a desaparecer. El presente estudio buscó identificar factores individuales y ambientales que pudieran predecir si se encontrará a un adulto mayor reportado como desaparecido. Se utilizaron modelos de aprendizaje automático supervisado basados en los datos abiertos de casos de personas desaparecidas de Colombia entre 1930 y junio de 2021 (n = 7.855). Se entrenaron algoritmos de clasificación para predecir si eventualmente se encontraría a un adulto mayor desaparecido. Los modelos de clasificación con mejor rendimiento en los datos de prueba fueron: Gradient Boosting Classifier y Light Gradient Boosting Machine, los cuales mostraron, respectivamente, un 10 % y un 9 % más AUC que un

modelo de referencia basado en el media del tiempo transcurrido desde la desaparición. Los rasgos que más contribuyeron a la clasificación fueron: el tiempo transcurrido desde la desaparición, el lugar donde ocurrió, la edad y sexo de la persona desaparecida. Los presentes resultados arrojan luz sobre el fenómeno social de la desaparición de personas y sientan las bases para la aplicación de modelos de aprendizaje automático en casos de personas mayores desaparecidas.

Palabras clave: Envejecimiento; Clasificación; Aprendizaje automático; Inteligencia artificial; Personas desaparecidas; Adultos mayores.

Abstract

Background. Person missingness is an enigmatic and frequent phenomenon that can bring about negative consequences for the missing person, their family, and society in general. Age-related cognitive changes and a higher vulnerability to dementia can increase the propensity of older adults to go missing. Thus, it is necessary to better understand the phenomenon of missingness in older adults. **Objective.** The present study sought to identify individual and environmental factors that might predict whether an older adult reported missing will be found. **Method.** Supervised machine learning models were used based on the missing person cases open data of Colombia between 1930 and June 2021 (*n* = 7,855). Classification algorithms were trained to predict

whether an older adult who went missing would eventually be found. **Results.** The classification models with the best performance in the test data were those based on gradient boosting. Particularly, the Gradient Boosting Classifier and the Light Gradient Boosting Machine algorithms showed, respectively, 10% and 9% greater area under the curve (AUC) of the receiver operating characteristic (ROC) curve than a data-driven, reference model based on the mean of the reported time elapsed since the missingness observed in the training data. The features with the greatest contribution to the classification were the time since the missingness, the place where it occurred, and the age and sex of the missing person. **Conclusion.** The present results shed light on the societal phenomenon of person missingness while setting the ground for the application of machine learning models in cases of missing older persons.

Keywords: Aging; Classification; Machine Learning; Artificial intelligence; Missing persons; Older adults

Introducción

La desaparición de personas es un fenómeno enigmático pero frecuente que puede tener consecuencias negativas para la persona desaparecida, sus familiares y la sociedad. Los adultos mayores (por ejemplo, los mayores de 60 años) pueden ser vulnerables a desaparecer. El envejecimiento puede tener un impacto negativo en las funciones cognitivas, como la atención, la memoria o el control cognitivo [1–3]. La edad también aumenta el riesgo de depresión, deterioro cognitivo o demencia [4]. Por ejemplo, los adultos mayores en las primeras etapas de la demencia pueden desaparecer mientras deambulan [5, 6], y en cualquier etapa de la demencia los adultos mayores pueden estar involucrados en uno o más incidentes de desaparición [7]. En otros casos, los adultos mayores con o sin síntomas depresivos pueden desaparecer 'voluntariamente' para planear o suicidarse [8]. Además, el maltrato a personas mayores, incluido el aislamiento social, la soledad o el abandono [9], también puede modificar el riesgo de desaparición de una persona mayor. Las consecuencias negativas sobre la salud mental o la integridad física [10] también pueden ser más graves en los adultos mayores porque pueden desorientarse más en el tiempo, el lugar o incluso en la persona mientras están desaparecidos. Una mayor desorientación, a su vez, disminuye la probabilidad de que una persona mayor desaparecida sea encontrada o pueda regresar a su hogar por sí misma. Además, las condiciones médicas crónicas que requieren múltiples medicamentos son más frecuentes en los

adultos mayores [11], lo que a su vez hace que encontrar a la persona mayor desaparecida sea aún más imperativo. Por lo tanto, una mejor comprensión de los factores que modifican la probabilidad de encontrar a un adulto mayor reportado como desaparecido puede arrojar luz sobre el fenómeno de la desaparición en general, pero también puede tener implicaciones prácticas para abordar el problema de manera más efectiva.

Numerosos factores individuales y ambientales pueden modificar la probabilidad de encontrar a un adulto mayor desaparecido [12], a través de las pistas y orientación que ofrecen a los investigadores de casos desaparecidos [13] y/o a la persona mayor desaparecida (por ejemplo, para ayudarla a regresar) . Por ejemplo, los mayores recursos cognitivos de una persona desaparecida o sus lazos sociales más estrechos podrían aumentar la probabilidad de que regrese si desapareciera involuntariamente. Además, un contexto ambiental más organizado en el que se produce la desaparición podría proporcionar mejores pistas al investigador que busca a la persona desaparecida (véase, por ejemplo, el trabajo experimental de [14] sobre el papel de la información espacial en la búsqueda), mientras que en el mismo tiempo puede ayudar a la persona desaparecida a encontrar el camino de regreso. Por lo tanto, el presente trabajo tuvo como objetivo predecir la probabilidad de encontrar a una persona mayor desaparecida e identificar los factores relevantes para esa predicción con base en modelos de aprendizaje automático supervisado.

El aprendizaje automático es una herramienta de inteligencia artificial que le permite a una computadora inferir las reglas que son necesarias para construir predicciones automáticamente [15, 16]. Las tareas de clasificación de aprendizaje automático son una herramienta adecuada [17, 18] para el estudio de fenómenos sociales y psicológicos complejos [19, 20], como los casos de personas desaparecidas. El trabajo anterior ha utilizado métodos de aprendizaje automático para investigar los perfiles de personas desaparecidas o para predecir la probabilidad de encontrarlas. En consecuencia, el trabajo pionero utilizó la minería de datos para elaborar reglas para predecir el resultado de los casos de personas desaparecidas y, por lo tanto, respaldar las intuiciones de los investigadores policiales involucrados en esos casos [21]. El trabajo reciente también ha propuesto utilizar modelos de aprendizaje automático durante la búsqueda de personas desaparecidas (por ejemplo, con reconocimiento facial [22] o fusión de datos multimodal basada en características [23]). Otros métodos están utilizando datos de dispositivos de seguimiento del sistema de posicionamiento global para intentar predecir ubicaciones típicas [24] o patrones de movilidad [25] de personas con demencia, que pueden tener un mayor riesgo de deambular y perderse, pero que aún no han desaparecido.

Un estudio reciente con una muestra de personas desaparecidas mostró un desempeño adecuado de modelos tales como K-vecinos más cercanos y árboles de decisión para predecir si una persona desaparecida se encuentra viva o muerta y si se

encuentra una persona desaparecida (independientemente de si está viva o muerta).) vs no encontrado, respectivamente [26]. Este estudio anterior se basó en datos sobre personas desaparecidas de todas las edades reportadas como desaparecidas en 2017. Otro estudio reciente sobre una muestra superpuesta usó el entorno de Waikato para el análisis de conocimiento y encontró perfiles que vinculan las causas de la desaparición (p. ej., desaparición 'voluntaria' vs. desaparición forzada) a determinados lugares y grupos de edad [27]. Sin embargo, a pesar de las condiciones particulares y la vulnerabilidad de los adultos mayores, ningún estudio, hasta donde sabemos, ha investigado el fenómeno de los adultos mayores desaparecidos en Colombia por causas diferentes a la desaparición forzada en los últimos 50 años.

En resumen, el presente estudio tuvo como objetivo identificar los factores individuales y ambientales que predicen si se encontrará a un adulto mayor desaparecido, mediante el uso de algoritmos de aprendizaje automático supervisado. Para ello, utilizamos datos abiertos proporcionados por el sistema de información de la Red de Desaparecidos y Cadáveres (Sistema de Información Red de Desaparecidos y Cadáveres, SIRDEC) del Instituto Nacional de Medicina Legal y Ciencias Forenses de Colombia. Nuestros objetivos específicos fueron (i) encontrar la probabilidad de encontrar a una persona mayor desaparecida, mediante el uso de algoritmos de clasificación y (ii) identificar qué características individuales o ambientales de las personas desaparecidas contribuyen a esa probabilidad, mediante el uso de aprendizaje automático interpretativo.

Materiales y método

Datos

El presente estudio utilizó los datos abiertos proporcionados por el SIRDEC del Instituto Nacional de Medicina Legal y Ciencias Forenses de Colombia (Instituto Nacional de Medicina Legal y Ciencias Forenses de Colombia) a través de la iniciativa Datos Abiertos del gobierno colombiano (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 datos se descargaron el 5 de agosto de 2021. La versión original de la base de datos incluía 162 401 entradas (es decir, ejemplos) de personas que desaparecieron en cualquier momento en el período de 1930 a junio de 2021. El presente estudio se realizó en tres fases: (i) limpieza de datos y selección de ejemplos y características relevantes, (ii) análisis descriptivos e (iii) identificación de modelos, evaluación e interpretación de modelos.

Preparación de datos y variables

En la primera fase, se excluyeron los ejemplos con información nula sobre las variables Edad (n = 202) y Fecha de extravío (n = 129). Esto se hizo así por dos razones. Primero, para asegurar que un ejemplo sí correspondía a un adulto mayor y, segundo, para asegurar la exactitud de la fecha de extravío. Los ejemplos cuya causa de desaparición fue “supuesta desaparición forzada” (n = 32.403) se excluyeron además

en función del objetivo del estudio. La razón para hacerlo fue que esta causa hace que sea más difícil encontrar patrones predictivos, ya que depende de factores posiblemente más complejos (por ejemplo, conflicto social y violencia), externos a la persona desaparecida. Luego de este paso, se aplicaron los siguientes criterios de exclusión: edad al momento de la desaparición menor de 60 años, estado actual “Encontrado muerto”, y país de la desaparición distinto a Colombia. Estos criterios dejaron 7.855 ejemplos válidos.

Las variables predictoras incluidas fueron la fecha y el lugar de la desaparición, como variables "ambientales" o extrínsecas, y la edad, el sexo, el estado civil, el nivel educativo y el factor de vulnerabilidad, como variables "individuales" o intrínsecas. Otras variables inicialmente disponibles, como 'país de nacimiento' o 'ascendencia racial', se excluyeron porque tenían el mismo valor en casi todos los ejemplos incluidos (es decir, "Colombia" y "mixto", respectivamente) y no se consideraron relevantes en la muestra actual. En la última parte de esta fase, algunas de las variables fueron transformadas para el paso de entrenamiento del modelo (Tabla 1), y luego se realizó un análisis descriptivo para cada variable, para identificar la distribución de datos, así como los valores faltantes.

Table 1. Variable transformation

variable original	Nueva variable	Categorías de la nueva variable = categorías de la variable original
Variable objetivo		
Status	Found	0 = “Aún vivo” 1 = “Encontrado vivo”
Explicativas		
Date	Elapsed time (in days)	Número de días hasta el 30 de julio de 2021 = Fecha de la pérdida
Marital status	Relationship	Current = “Convivencia con pareja”, “Casado” Past = “Divididos”, “Divorciados”, “Viudos” None = “Soltero”
Education level	Education (in years)	0.0 = “None” 2.5 = “Educación inicial y preescolar” 5.0 = "Escuela primaria" 7.5 = "Escuela media" 10.0 = “Bachillerato” 12.5 = “Grado asociado” 15.0 = “Universitario” 17.5 = “Especialización o maestría o equivalente” 20.0 = “Doctorado o equivalente”

Vulnerability factor	Vulnerability	No = “None” Yes = Todos los valores excepto “Sin información”
Municipality and department of missingness	Municipality (inhabitants)*	Total de habitantes (2015 – 2018) obtenido de Wikipedia (https://es.wikipedia.org/wiki/Municipios_de_Colombia) incluyendo apéndices (e.g., https://es.wikipedia.org/wiki/Anexo:Municipios_de_Huila)

* Calculado usando web-scraping: [https://github.com/virtualmarioe/Web_scraping_tutorial]

Preprocesamiento y modelado de datos

La tercera fase comprendía el preprocesamiento y el modelado. En el preprocesamiento, primero, los datos se dividieron en conjuntos de entrenamiento y prueba, utilizando el 80 % (n = 6284) y el 20 % (n = 1571) de los datos, respectivamente. Los datos se dividieron aleatoriamente usando la función train_test_split, estratificando por clase (es decir, "Aún vivo" y "Encontrado vivo"). Este paso aseguró que tanto los conjuntos de datos de entrenamiento como los de prueba tuvieran la misma representación de clase, ya que el 65,8 % (n = 5166) de los ejemplos tenían la etiqueta "Aún vivo" y el 34,2 % (n = 2689) la etiqueta "Encontrado vivo" en todo el marco de datos. A continuación, se imputaron los valores faltantes en los conjuntos de entrenamiento y prueba, usando la media correspondiente de las variables numéricas con valores faltantes (es decir, Educación y Municipio) en los datos

de entrenamiento. Asimismo, los valores faltantes se imputaron en los conjuntos de entrenamiento y prueba, utilizando el modo correspondiente de las variables categóricas con valores faltantes (es decir, Vulnerabilidad y Relación) en los datos de entrenamiento. La imputación se realizó a través de la función *SimpleImputer*, se ajustó a los datos de entrenamiento y luego se aplicó a los conjuntos de datos de entrenamiento y prueba.

A continuación, se propuso un modelo de referencia (o base) simple, basado (únicamente) en los datos de entrenamiento. Este modelo basado en reglas se usó simplemente para juzgar el rendimiento de los modelos de aprendizaje automático. Además, las variables numéricas y categóricas se transformaron con escala estándar y codificación one-hot, respectivamente, para tener solo características numéricas como entrada para los modelos. De igual manera, la variable resultado se ajustó con la función Label Encoder. Nuevamente, el ajuste variable se realizó ajustando solo los datos de entrenamiento (es decir, para evitar la fuga de datos) y luego se aplicó a ambos conjuntos de datos (entrenamiento y prueba).

La clase "Aún vivo" tenía casi el doble de ejemplos en la clase "Encontrado vivo" (es decir, 65,8 % frente a 34,2 % tanto en los datos de entrenamiento como de prueba). Por lo tanto, entrenamos los modelos con datos remuestreados en el conjunto de entrenamiento solo como un medio para evitar que los modelos estén sesgados hacia la clase mayoritaria. Por lo tanto, se logró una distribución equilibrada (es decir, 50/50)

de clases en los datos de entrenamiento a través de (a) técnica de sobremuestreo de minorías sintéticas (SMOTE) (ntrain(1) = ntrain(2) = 4,133) y (b) submuestreo (ntrain(1) = ntrain(2) = 2,151). Para completar y transparencia, los resultados también se presentan usando todos los datos de entrenamiento disponibles durante el entrenamiento del modelo (es decir, sin remuestreo; Tabla 2).

En la parte de modelado, primero se realizó un análisis global de los algoritmos de clasificación (Figura S1) con una validación cruzada estratificada de 10 veces (variable de resultado, "Encontrado vivo": 0 = "no", 1 = "sí"). A continuación, se seleccionaron los tres modelos con las puntuaciones de precisión más altas (es decir, número de predicciones correctas/número total de predicciones) para cada estrategia de remuestreo, a partir de las cuales se examinaron sus matrices de confusión. Otras métricas de rendimiento, como el recuerdo (es decir, la identificación de casos positivos verdaderos de todos los casos positivos posibles), la precisión (es decir, la identificación de casos positivos verdaderos de todos los casos identificados como positivos), el área bajo la curva (AUC) de la También se evaluaron la curva característica del operador de recepción (ROC) (es decir, la capacidad de distinguir entre clases positivas y negativas) y la puntuación F1 (es decir, sensibilidad y especificidad de ponderación media armónica). La extracción de la importancia de las características para la interpretación de las predicciones del modelo se realizó con los valores SHApley Additive exPlanation (SHAP) [28]. Las tres fases analíticas se

realizaron en Python (v. 3.6) con la biblioteca PyCaret (https://pycaret.org/) (v. 2.3.6) y Scikit Learn (v. 1.0.2) (https://scikit-learn.org/stable/) [29].

Disponibilidad de datos

Los datos y el código en los que se basan los resultados del presente estudio están disponibles abiertamente y se pueden encontrar en [https://osf.io/ysmf5/?view_only=25257cab6d0943439cd1b5b69cdbc5a9].

Resultados

Análisis descriptivo

La distribución de ejemplos "Encontrado vivo" y "aún vivo" a lo largo de meses y años se presenta en la Figura 1. En general, los casos "aún vivo" aparecen escasos antes del año 1980, y los casos "Encontrado vivo" aparecen escasos antes de 2000. En toda la muestra , la edad media de los ejemplos con estado "Encontrado vivo" fue de 71,35 ± 8,36 años (vs. 71,45 ± 9,91 años de "Aún vivo") (Figura 2) y la educación media fue de 5,12 ± 3,53 años (vs. 4,85 ± 3,39 años de “Siguen desaparecidos”). La mayoría de los ejemplos fueron del sexo masculino (72,8% "Encontrado vivo" vs. 83% “aún vivo”) y corresponden a casos de adultos mayores sin factor de vulnerabilidad evidente (74,3% “Encontrado vivo” vs. 71,7% “aún vivo”) y con una relación actual (40,2 % "Encontrado vivo" frente a 49,2 % “Todavía desaparecido”) en el momento del informe

de desaparición. Casi la mitad de los casos de desapariciones ocurrieron en municipios con población menor a 1 millón de habitantes, casi el 36% ocurrieron solo en la ciudad capital (con aprox. 8 millones de habitantes), y una mayor proporción de casos "Encontrado vivo" ocurrieron en municipios con población por encima de los 2 millones de habitantes (Figura 3). La mayoría de los casos se informaron hace menos de 5000 días (es decir, 14 años aproximadamente), siendo este número el límite superior para casi todos los casos con el estado "Encontrado vivo" (Figura 4).

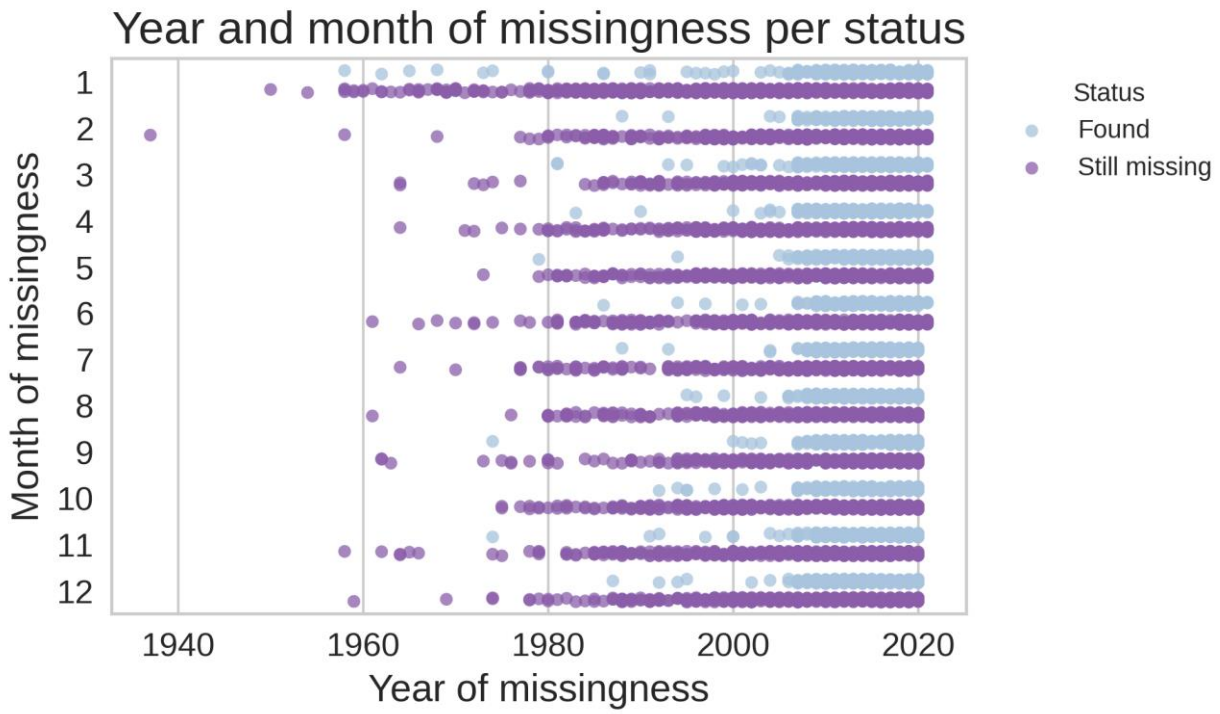

Figura 1. Gráfico de franjas de la distribución por clase a lo largo del tiempo. El mes y el año del informe de faltantes se muestran para cada clase (es decir, 'Encontrado vivo' y 'Aún vivo') en todo el marco de datos.

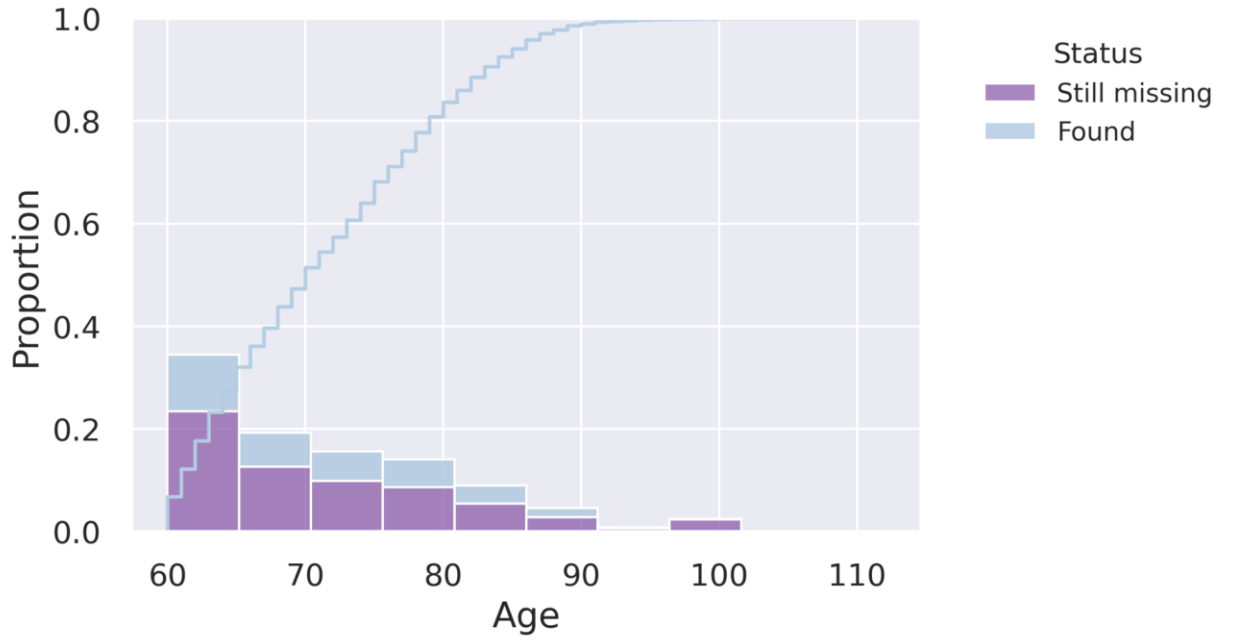

Figura 2. Histograma de edad (en años) por el estado faltante de todo el marco de datos (n = 7855). La distribución por edades fue similar en ambos grupos de estatus. La línea azul clara sobre las barras del histograma representa la función de distribución acumulativa empírica o la proporción de ejemplos que están debajo de cada valor único en el conjunto de datos.

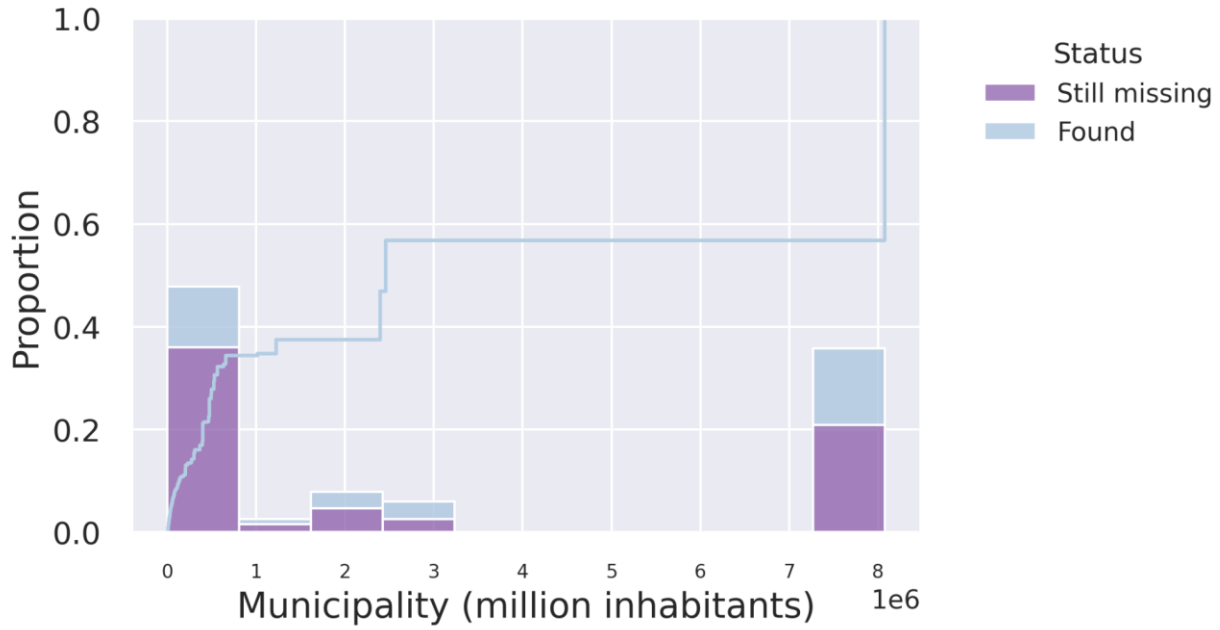

Figura 3. Histograma del tamaño del municipio (en millones de habitantes) del lugar donde se reportó la ocurrencia del caso desaparecido por estado de desaparición (“Aún vivo” o “Encontrado vivo”). La línea celeste sobre las barras indica la proporción acumulada de los ejemplos con estado “Encontrado vivo”: la mayor proporción de casos con estado “Encontrado vivo” se observa en municipios con población superior a dos millones de habitantes.

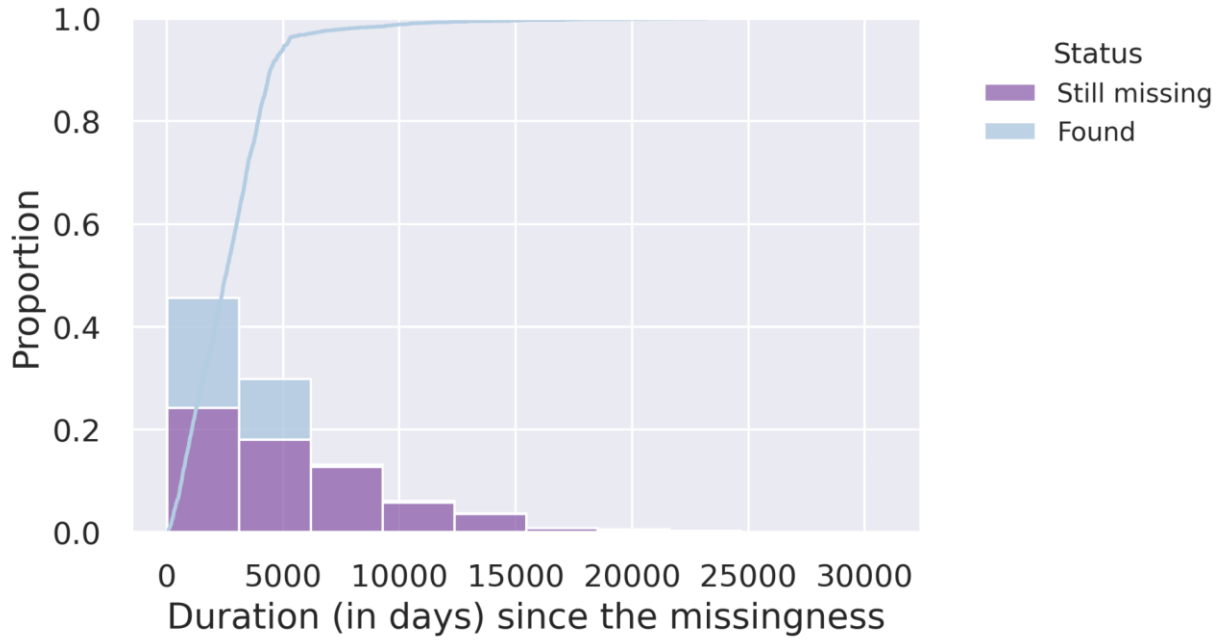

Figura 4. Histograma de fecha de extravío (en número de días desde el reporte hasta el 30 de julio de 2021) por estado de extravío (“Aún vivo” o “Encontrado vivo”). La línea azul claro sobre las barras indica la proporción acumulada de ejemplos con el estado "Encontrado vivo": más del 90 % de los casos con el estado "Encontrado vivo" tienen una fecha de notificación inferior a 5000 días o 14 años aproximadamente (es decir, desaparecieron en 2007 o posterior). Tenga en cuenta que esta variable representa el contexto temporal de la desaparición (es decir, el cuándo) y no la duración real de la desaparición para los casos "Encontrado vivo", que se incluye en los datos.

Modelo base

Siguiendo las ideas del análisis descriptivo, se formuló un modelo base como modelo de referencia. Este modelo solo sirvió para permitirnos juzgar el rendimiento de los modelos

de aprendizaje automático, pero no para sacar ninguna conclusión. El modelo base fue la media del tiempo transcurrido (en días) desde el informe de desaparición, cuál fue la regla predictiva para el resultado, es decir, si se encontrará a la persona mayor desaparecida. Tenga en cuenta que elegimos el tiempo medio transcurrido como regla debido a su simplicidad y porque se puede estimar fácilmente a partir de los datos existentes. Esta regla (4474,8 días en los datos actuales) se calculó sólo en el conjunto de entrenamiento y luego se aplicó al conjunto de prueba, lo que arrojó una precisión del 63% (Tabla 2). Por lo tanto, el rendimiento del modelo de aprendizaje automático se comparó y se juzgó con respecto a esta precisión del 63 % de referencia.

Tabla 2. Rendimiento medio en los datos de prueba (con validación cruzada de 10 veces) de los tres mejores modelos con y sin corrección de desequilibrio de clase (ordenados según la precisión).

Modelo	Accuracy	AUC	Recall	Precision	F1
Without class imbalance fix in the training data					
Gradient Boosting Classifier	0.72	0.79	0.53	0.60	0.56
AdaBoost Classifier	0.71	0.77	0.55	0.59	0.56
Light Gradient Boosting Machine	0.71	0.78	0.52	0.59	0.55

Oversampling the minority class in the training data with SMOTE					
Light Gradient Boosting Machine	0.71	0.78	0.67	0.56	0.61
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
Undersampling the majority class in the training data with RandomUnderSampler					
Gradient Boosting Classifier	0.68	0.79	0.85	0.52	0.65
Light Gradient Boosting Machine	0.68	0.77	0.79	0.52	0.63
Random Forest Classifier	0.68	0.76	0.73	0.52	0.61
Without machine learning: Rule-based model [a]					
Reference or base model	0.63	0.69	0.89	0.48	0.62

[a] *Tiempo medio transcurrido (en días) desde la desaparición (4474,8 días), independientemente de la duración de la desaparición en los casos encontrados, que se desconoce en los datos actuales*

Modelos de aprendizaje automático

Los tres "mejores" modelos para cada estrategia de corrección de desequilibrio de clase se enumeran en la Tabla 2 (consulte la Tabla complementaria S1 para obtener un informe de las métricas de todos los modelos sin usar la corrección de desequilibrio de clase durante el entrenamiento del modelo). El rendimiento fue similar entre ellos en

todas las métricas en las estrategias de remuestreo de datos de entrenamiento (incluido el no remuestreo). Sin embargo, Recall mejoró sustancialmente cuando se usó submuestreo en los datos de entrenamiento. Tanto el clasificador de aumento de gradiente (GBC) como la máquina de aumento de gradiente de luz (LGBM) se encontraban entre los mejores modelos, independientemente de si se solucionó o no el desequilibrio de clase.

Examinamos con mayor detalle el GBC entrenado con datos de entrenamiento submuestreados, ya que tanto con SMOTE como sin corrección de desequilibrio, la clase minoritaria (es decir, "Encontrado vivo") fue penalizada en la mayoría de las métricas, incluso en los modelos más precisos (consulte las Figuras complementarias S1 y S2).). Como se puede observar en la matriz de confusión (Figura 5), el 17 % de los ejemplos fueron falsos negativos (es decir, casos "Encontrado vivo" que se predijo que “Aún vivo”), mientras que el 41 % de los ejemplos fueron falsos positivos (es decir, , "Aún vivo" casos que se predijo que serían "Encontrado vivo"). En particular, la tasa de falsos positivos representa una mejora sustancial con respecto al modelo de referencia o base, en el que este porcentaje se encontraba en el nivel del azar (falsos positivos) (Figura complementaria S3). Además, la puntuación AUC aumentó al menos un 7 % con respecto al modelo de referencia en todos los mejores modelos en todas las estrategias de remuestreo (Tabla 2 y Figura 6). El AUC fue similar en los mejores modelos de aprendizaje automático (es decir, 0,76 – 0,79; consulte también las Figuras

complementarias S5 y S6 para comparar). Finalmente, el modelo GBC que usó submuestreo de los datos de entrenamiento mostró una métrica de recuperación más alta (es decir, 0.83) y una puntuación F1 más alta (es decir, 0.63) en la clase "Encontrado vivo" (es decir, la clase de interés; Figura complementaria S4) en comparación con el modelo LGBM entrenado con SMOTE (Figura complementaria S2) (recuerdo: 0,65; puntaje F1: 0,60) y el modelo GBC entrenado sin volver a muestrear los datos de entrenamiento (recuerdo: 0,52; puntaje F1: 0,55; Figura complementaria S2).

Figura 5. Matriz de confusión del modelo clasificador de aumento de gradiente con submuestreo de los datos de entrenamiento. Las puntuaciones de clasificación se normalizan por fila. Para la clase "Encontrado vivo", el 83% de los casos están correctamente clasificados, mientras que el 59% de los casos están correctamente clasificados para la clase "Aún vivo".

Figura 6. Curva característica operativa del receptor (ROC) para el clasificador de aumento de gradiente (GBC). La tasa de verdaderos positivos es mayor que la tasa de falsos positivos. Los valores de AUC fueron similares en todos los modelos de aprendizaje automático (que se muestran en la Tabla 2). La línea negra punteada representa un clasificador 'ficticio' con AUC = 0,50.

Funciones relevantes para la predicción en casos de personas mayores desaparecidas

El segundo objetivo del presente estudio fue identificar los factores que determinan si

un adulto mayor desaparecido en Colombia será encontrado más tarde. En consecuencia, examinamos la importancia de la característica, es decir, la contribución relativa de la característica a la predicción en el modelo GBC (Figura 7). Las características identificadas fueron el número de días transcurridos desde la denuncia de la desaparición, el tamaño del municipio (en número de habitantes) donde ocurrió la desaparición, el sexo de la persona desaparecida y la edad de la persona desaparecida al momento de la denuncia. En la Figura complementaria S7 se pueden observar algunos ejemplos de los valores de estas variables, así como de las predicciones específicas en el conjunto de datos de prueba.

Para identificar las características que más contribuyeron a la predicción del modelo, examinamos la importancia de la característica como una función de los valores de explicación del aditivo SHapley (SHAP) (Figura 7) para el modelo GBC con datos de entrenamiento submuestreados. Un mayor tiempo transcurrido (en días) desde el reporte de la desaparición, un municipio pequeño (es decir, con una población relativamente menor), ser hombre y una edad más avanzada de la persona desaparecida se asociaron con una menor probabilidad de que un adulto mayor desaparecido ser encontrado más tarde.

Figura 7. Importancia de la característica del modelo GBC con datos de entrenamiento submuestreados. Las funciones utilizadas para la predicción se muestran por orden de relevancia en el eje Y y los valores SHAP (SHapley Additive exPlanations) se muestran

en el eje X, donde los valores negativos representan la etiqueta "Aún vivo" y los valores positivos, el etiqueta "Encontrado vivo". Cada punto es un ejemplo del conjunto de datos de entrenamiento. Los códigos de escala de colores para el valor de un ejemplo en particular: puntos azules, valores bajos; puntos morados, valores intermedios; puntos rojos, valores altos.

Impacto potencial de los cambios sociales durante 90 años en los casos de personas desaparecidas

Si bien la mayoría (es decir, el 83,8 %) de nuestros ejemplos fueron reportados como desaparecidos en 2000 y más tarde, nuestros datos abarcan casos de personas mayores desaparecidas desde 1930 hasta mediados de 2021 (Figura 1). Muchos cambios sociales han ocurrido durante estos 90 años, y las nuevas tecnologías ciertamente han permitido mejorar la búsqueda, el reporte y el registro de casos de personas desaparecidas. Por lo tanto, post hoc, restringimos los ejemplos solo a los de los últimos 20,5 años (n = 6582; "Encontrado vivo": 2638; "Aún vivo": 3944), para reducir el impacto potencial de los cambios sociales y tecnológicos en la formación de modelos y actuación. Por lo tanto, repetimos el entrenamiento del modelo submuestreando los datos de entrenamiento de acuerdo con lo descrito en las dos secciones anteriores. La Tabla 3 enumera los modelos más precisos. De acuerdo con los resultados de datos de 1930-2021', GBC superó al modelo de referencia en todas

las métricas. Las métricas del modelo de aprendizaje automático se mantuvieron sólidas y fueron similares a las obtenidas sin restringir los datos a los años más recientes (Tabla 2). Por el contrario, el modelo basado en reglas, que depende en gran medida del tiempo transcurrido, disminuyó notablemente su rendimiento. Finalmente, la importancia de la característica también fue comparable a la que se utilizó con los datos de 1930-2021 (Figura 8).

Tabla 3. Desempeño promedio (con validación cruzada de 10 veces) en los datos de prueba (restringido a faltantes en 2000 y posteriores) de los modelos con la mayor precisión

Modelo	Accuracy	AUC	Recall	Precision	F1
Undersampling the majority class in the training data with RandomUnderSampler					
Random Forest Classifier	0.64	0.71	0.69	0.54	0.61
Light Gradient Boosting Machine	0.64	0.72	0.71	0.54	0.61
Gradient Boosting Classifier	0.64	0.73	0.80	0.53	0.64
AdaBoost Classifier	0.63	0.72	0.80	0.52	0.63
Without machine learning: Rule-based model [a]					
Reference or base model	0.57	0.59	0.65	0.48	0.55

[a] Tiempo medio transcurrido desde la perdida: 3110.2 días u 8.5 años.

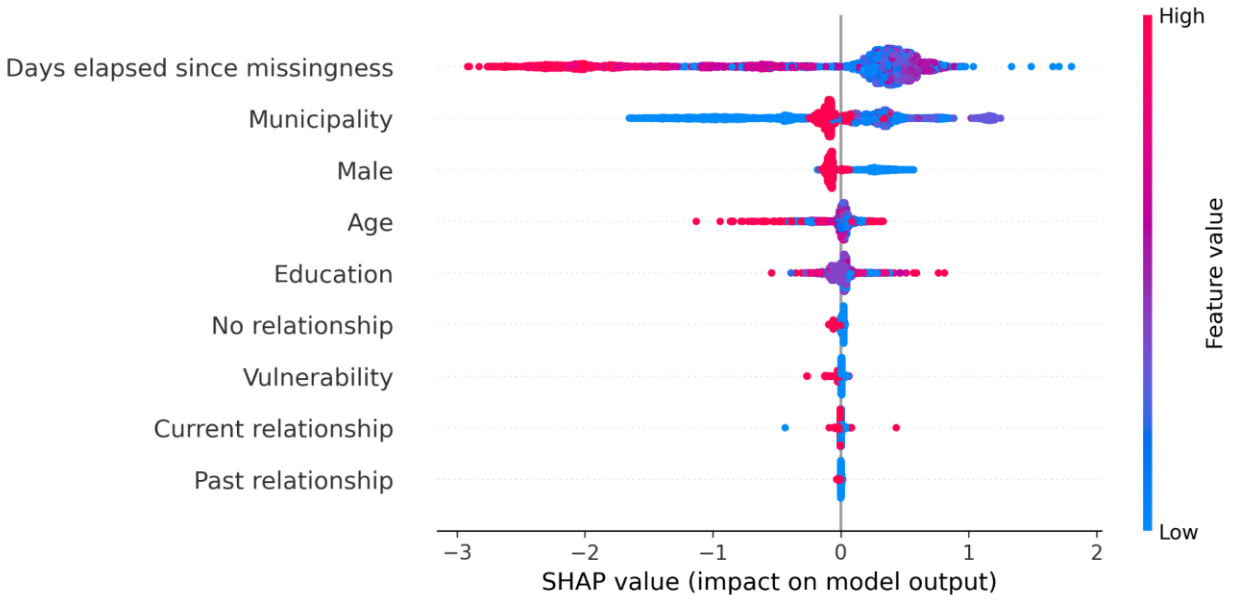

Figura 8. Importancia característica del modelo GBC en datos con faltantes en el año 2000 y posteriores. Las funciones utilizadas para la predicción se muestran por orden de relevancia en el eje Y y los valores SHAP (SHapley Additive exPlanations) se muestran en el eje X, donde los valores negativos representan la etiqueta "Aún vivo" y los valores positivos, el etiqueta "Encontrado vivo". Tenga en cuenta que la importancia de la característica puede diferir ligeramente de la observada cuando se utilizan los datos de '1930-2021' para el entrenamiento del modelo (Figura 7). Cada punto es un ejemplo del conjunto de datos de entrenamiento. Los códigos de escala de colores para el valor de un ejemplo en particular: puntos azules, valores bajos; puntos morados, valores intermedios; puntos rojos, valores altos.

Discusión

El presente estudio buscó identificar los factores individuales y ambientales que predicen si se encontrará a un adulto mayor desaparecido, mediante el uso de aprendizaje automático supervisado. Los resultados mostraron que los mejores modelos para este propósito eran los basados en conjuntos y, más específicamente, en el aumento de gradiente; en particular, la máquina potenciadora de gradiente ligero (LGBM) y el clasificador del potencializador del gradiente (GBC). El error de clasificación de los modelos de aprendizaje automático (es decir, entre el 28 % y el 32 %) estuvo por debajo del nivel de error de un modelo base (es decir, el 37 %) que usó el tiempo medio transcurrido (en días) desde el informe de ausencia en el datos de entrenamiento como regla de predicción. Este hallazgo indica que los modelos de aprendizaje automático pueden informarnos sobre los factores que predicen el resultado de los casos de personas mayores pérdidas y, al mismo tiempo, generar una predicción para cada caso individual. Los factores identificados como cruciales para predecir que una persona desaparecida será posteriormente encontrada fueron el menor tiempo transcurrido desde el reporte de la desaparición, un municipio de tamaño relativamente mediano donde ocurre la desaparición, el sexo femenino y una edad menos avanzada de la persona desaparecida. El rendimiento del modelo de aprendizaje automático fue sólido incluso cuando solo se usaron datos de los últimos 20.5 años para el entrenamiento y las pruebas del modelo. Juntos, los hallazgos

presentes brindan información sobre el complejo fenómeno social de la desaparición de personas en adultos mayores y potencialmente tienen implicaciones prácticas.

Los modelos de clasificación más precisos en el estudio actual fueron modelos basados en conjuntos de árboles de decisión, por ejemplo, potencializador del gradiente [30, 31] y Random Forest [32]. Este resultado se alinea bien con informes anteriores [21, 26]. La mayoría de los clasificadores (p. ej., K-Neighbors, SVM con kernel lineal, análisis discriminante lineal; consulte también la Tabla S1) se desempeñaron bien en la mayoría de las métricas. Ejemplos notables en el estudio actual fueron el GBC y el LGBM, que tuvieron las métricas de rendimiento más altas, independientemente de si el desequilibrio de clase se soluciona o no durante el entrenamiento del modelo. GBC es, en términos simples, un conjunto de modelos iterativos, en el que cada vez se entrena un modelo nuevo y débil teniendo en cuenta el error previamente aprendido del conjunto (ver, por ejemplo, [33]). LGBM es una implementación especial del algoritmo de árbol de decisión de aumento de gradiente [34]. En el presente estudio, un modelo GBC entrenado con datos balanceados a través del submuestreo de la clase dominante (es decir, "Aún vivo") nos permitió maximizar la métrica de recuerdo en ambas clases con respecto al modelo de referencia. Este resultado implica que GBC redujo la tasa de falsos positivos (es decir, la predicción de que un caso es "Encontrado vivo" cuando, en realidad, "Aún vivo") del 50% al 41% en comparación con el modelo de referencia, como se refleja en un mayor

AUC de la curva ROC (es decir, 79% de GBC vs. 69% del modelo de referencia). En términos prácticos, este resultado significa que nuestro modelo de aprendizaje automático puede predecir correctamente al menos un caso de persona desaparecida más de cada diez casos, en comparación con un modelo basado en datos informados (Figura 5 y Figura S3).

Para poner esos resultados en perspectiva, en primer lugar, sin un modelo difícilmente se puede generar una probabilidad confiable para el resultado de un caso de persona mayor desaparecida, o dicha probabilidad se basará únicamente en la intuición del investigador del caso desaparecido. En segundo lugar, con la referencia actual, el modelo basado en datos, solo el tiempo transcurrido desde la ausencia informa la predicción (es decir, por encima o por debajo de ~12 años). Aquí vale la pena mencionar que nuestro modelo de referencia (o base) basado en datos es congruente con informes empíricos sobre muestras más jóvenes de desaparición forzada en Colombia, con un tiempo promedio transcurrido de 13,38 ± 6,88 años [35]. El uso del tiempo medio transcurrido desde el informe de faltantes como regla implica que el modelo de referencia es principalmente útil como modelo explicativo pero menos útil como modelo predictivo, es decir, para los nuevos casos, todos los cuales tendrán inherentemente un tiempo transcurrido desde el ausencia menor de 12 años. Sin embargo, el valor del modelo base radica en que proporciona una línea de base significativa para comparar los modelos de aprendizaje automático. Por último, y en

marcado contraste con las dos opciones anteriores, con el modelo de aprendizaje automático identificado en el estudio actual, se pueden generar predicciones individuales sobre nuevos casos de personas desaparecidas. Este resultado representa un paso significativo para proporcionar un soporte sólido basado en computación [19] para la investigación de personas mayores desaparecidas y para el estudio de la desaparición de personas como un fenómeno social desde un enfoque cuantitativo y flexible [36]. En el futuro, se podrían dedicar algunos esfuerzos a entrenar y probar modelos más complejos, por ejemplo, aquellos basados en redes neuronales. Sin embargo, estos modelos tienden a funcionar de manera subóptima con datos tabulares [37] y es posible que no se generalicen bien [38].

Nuestro estudio también identificó las características que eran críticas para la predicción del resultado de ausencia. Como era de esperar, tanto los factores intrínsecos como los extrínsecos resultaron cruciales. Específicamente, la edad de la persona desaparecida, que se relaciona con el estado cognitivo [1–4, 39] o de salud global [11] de la persona, o el sexo de la persona desaparecida, que se relaciona con el motivo de la desaparición [40] o el tipo de comportamientos en que se involucra la persona durante la ausencia, fueron importantes. Asimismo, la fecha de la desaparición o el tamaño del municipio en el que se produjo la desaparición fueron relevantes, ya que están indirectamente asociados a la estructura y organización del medio físico y social que rodea a la desaparición. Por un lado, estos factores temporales y de lugar

muy probablemente reflejan el cambio social a lo largo de la segunda mitad del siglo XX y principios del siglo XXI (por ejemplo, en términos de infraestructura, tecnología, comunicaciones, crecimiento de la población y organización social). . Por otro lado, también pueden reflejar el creciente reconocimiento de las personas desaparecidas como un problema social común y la correspondiente promulgación y perfeccionamiento del registro y búsqueda de personas desaparecidas en Colombia. En general, estos resultados se prestan a futuras evaluaciones basadas en humanos y/o funcionales como otro medio para juzgar el rendimiento [41] de los modelos identificados en el presente estudio.

Contrariamente a nuestras expectativas, otros factores intrínsecos no parecen contribuir significativamente a la predicción. Estos factores fueron la vulnerabilidad, el estado civil y el nivel educativo de la persona mayor desaparecida. Una posible explicación de estos hallazgos negativos es la variabilidad de datos relativamente baja en estas características, además de la alta proporción de valores que les faltaban. Por lo tanto, en el futuro, la cuantificación de estas variables podría ayudar a dilucidar si tienen un impacto en la probabilidad de encontrar a la persona desaparecida. Ejemplos particulares a este respecto son registrar el número de personas con las que vivía la persona desaparecida; el número de factores de vulnerabilidad (p. ej., médicos, sociales, cognitivos) de la persona desaparecida; el número de años de educación de la persona desaparecida; el número de incidentes desaparecidos anteriores, si los

hubiere; o un grado de 'cercanía' dependiendo de quién informe la desaparición. Tres dimensiones del comportamiento pueden tipificar a una persona adulta desaparecida: disfuncional (es decir, problemas mentales, incluida la demencia [7]), escape (es decir, personas que deciden o se ven obligadas a desaparecer para ganar independencia o huir de las dificultades) y no intencional (es decir, , bajo la influencia de otros o como consecuencia de un accidente o problema de comunicación con sus allegados) [42]. Las tipologías que más caracterizan a los adultos mayores (es decir, mayores de 60 años) son disfuncionales y de escape [42]. Esta particularidad, sumada a la multiplicidad de circunstancias ambientales asociadas a la desaparición, implica que las consecuencias de la desaparición pueden impactar no solo a la persona desaparecida sino también a aquellos relacionados directa o indirectamente con ella [43]. Por ejemplo, en muchos casos, a los familiares les resulta difícil hacer el duelo, incluso muchos años después de la desaparición de su familiar [35]. En este contexto, las ideas del presente estudio podrían tener implicaciones prácticas tanto para el grupo de trabajo que se ocupa de los casos de personas desaparecidas como para el trabajo psicosocial con la familia de un adulto mayor desaparecido. En particular, se puede generar una mayor conciencia social sobre el resultado de la ausencia de los ancianos, especialmente los hombres (p. ej., mediante una amplia implementación de estrategias de identificación y reorientación, [44]). Del mismo modo, se pueden buscar mejoras específicas en los municipios más pequeños en los grupos de trabajo de personas

desaparecidas. Además, los profesionales psicosociales pueden utilizar la predicción de resultados en un caso específico para tomar mejores decisiones basadas en datos que les ayuden a adaptar su asesoramiento, por ejemplo, enfatizando las estrategias de afrontamiento que pueden ser más relevantes para ese caso específico.

Los presentes hallazgos deben ser considerados teniendo en cuenta algunas limitaciones. En primer lugar, los datos presentes no se recopilaron con fines de investigación científica y, por lo tanto, no incluyen todos los detalles relevantes para la teoría o la profundidad de la información o pueden no ser precisos. En segundo lugar, hubo una gran cantidad de valores faltantes, que manejamos a través de métodos de imputación simple. Por lo tanto, podría haber cierto grado de incertidumbre en las predicciones debido a esos aspectos. En tercer lugar, y como consecuencia de ello, los datos eran ruidosos y es posible que no hayan permitido un mejor rendimiento del modelo. Sin embargo, es importante tener en cuenta que los casos de personas desaparecidas son un fenómeno social intrínsecamente complejo. Más importante aún, cada punto porcentual obtenido con cualquier modelo dado se traduce en un caso de persona desaparecida que se predice correctamente, lo que en última instancia justifica el uso del modelo y su mejora. Finalmente, los estudios futuros deben determinar si los presentes hallazgos y conclusiones se generalizan también a casos de personas desaparecidas que involucran a adultos jóvenes o niños o en los que hubo desaparición forzada o el desenlace fatal, o a casos de personas mayores

desaparecidas en otros países. Sin embargo, a pesar de sus limitaciones, el presente estudio arrojó elementos para una mejor comprensión de los factores que predicen que un adulto mayor desaparecido en Colombia será posteriormente encontrado y sentó un precedente en términos de algoritmos de inteligencia artificial que pueden ser idóneos para abordar el problema de la predicción de resultados en casos de adultos mayores desaparecidos.

Conclusión

El presente estudio identificó los factores individuales (como la edad y el sexo) y ambientales (como el tiempo transcurrido y el tamaño del lugar de la desaparición) que predicen si se encontrará a un adulto mayor desaparecido, mediante el uso de un modelo de aprendizaje automático supervisado basado en conjuntos. Los presentes hallazgos sugieren que hay factores intrínsecos y extrínsecos en juego, todos los cuales pueden influir en la predicción del resultado. Estos factores son el estado cognitivo de la persona desaparecida antes o durante la desaparición, el tipo de conductas que adopta la persona durante la desaparición y la estructura y organización del entorno físico y social que la rodea. Además, este modelo de aprendizaje automático no solo redujo el error del modelo basado en datos de referencia en un 5 % y aumentó la discriminación de tasa positiva (es decir, la curva AUC-ROC) en un 10 %, sino que también nos permitió generar predicciones individuales para nuevos , casos no vistos. En general, el presente trabajo tiene implicaciones prácticas para los casos

de personas mayores desaparecidas, ya que puede ayudar a informar la decisión de los profesionales involucrados tanto en la búsqueda de personas mayores desaparecidas como en el trabajo psicosocial para apoyar a los familiares de la persona desaparecida.

Referencias

1. 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
2. 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
3. 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
4. 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
5. 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

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
7. Rowe, M., Houston, A., Molinari, V., Bulat, T., Bowen, M. E., Spring, H., … McKenzie, B. (2015). The Concept of Missing Incidents in Persons with Dementia. *Healthcare*, *3*(4), 1121–1132. https://doi.org/10.3390/healthcare3041121
8. Vargas Rodríguez, P. (2010, March 1). *Tras las huellas de los desaparecidos "voluntarios" en Bogotá* (bachelorThesis). *instname:Universidad del Rosario*. Universidad del Rosario. Retrieved from https://repository.urosario.edu.co/handle/10336/1778
9. World Health Organization, N. D. and M. H. C., & (INPEA), I. N. for the P. of E. A. (2002). Missing voices : views of older persons on elder abuse. World Health Organization. Retrieved from https://apps.who.int/iris/handle/10665/67371
10. LAI, C. K. Y., CHUNG, J. C. C., WONG, T. K. S., FAULKNER, L. W., NG, L., & LAU, L. K. P. (2012). MISSING OLDER PERSONS WITH DEMENTIA — A HONG KONG VIEW. *The Hong Kong Journal of Social Work*. https://doi.org/10.1142/S0219246203000214
11. Hayes, B. D., Klein-Schwartz, W., & Barrueto, F. (2007). Polypharmacy and the Geriatric Patient. *Clinics in Geriatric Medicine*, *23*(2), 371–390. https://doi.org/10.1016/j.cger.2007.01.002
12. 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. Retrieved from https://ufv.ca/media/assets/ccjr/reports-and-publications/Missing_Persons.pdf

13. 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
14. Moore, K. N., Lampinen, J. M., & Provenzano, A. C. (2016). The Role of Temporal and Spatial Information Cues in Locating Missing Persons. *Applied Cognitive Psychology*, *30*(4), 514–525. https://doi.org/10.1002/acp.3242
15. 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.
16. 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). Singapore: Springer. https://doi.org/10.1007/978-981-13-7403-6_11
17. 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
18. 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
19. Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. https://doi.org/10.1177/0002716215570279
20. Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology:

Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

21. 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). Berlin, Heidelberg: Springer. https://doi.org/10.1007/11011620_19
22. 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, Cartagena de Indias D.T y D.C.
23. Solaiman, K. M. A., Sun, T., Nesen, A., Bhargava, B., & Stonebraker, M. (2022). Applying Machine Learning and Data Fusion to the “Missing Person” Problem. https://doi.org/10.36227/techrxiv.16556121.v2
24. Wojtusiak, J., & Mogharab Nia, R. (2021). Location prediction using GPS trackers: Can machine learning help locate the missing people with dementia? *Internet of Things*, *13*, 100035. https://doi.org/10.1016/j.iot.2019.01.002
25. Bayat, S., & Mihailidis, A. (2021). Outdoor life in dementia: How predictable are people with dementia in their mobility? *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, *13*(1), e12187. https://doi.org/10.1002/dad2.12187
26. 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
27. 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.

28. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.
30. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
31. Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504
32. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
33. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. Retrieved from https://www.frontiersin.org/article/10.3389/fnbot.2013.00021
34. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information*

Processing Systems (Vol. 30). Curran Associates, Inc.

35. 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
36. Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921
37. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep Neural Networks and Tabular Data: A Survey. *arXiv:2110.01889 [cs]*. Retrieved from http://arxiv.org/abs/2110.01889
38. Blackmore, K., & Bossomaier, T. (2002). Soft computing methodologies for mining missing person data: Australia-Japan Joint Workshop on Intelligent and Evolutionary Systems. In N. Namatame (Ed.), *Sixth Australia-Japan Joint Workshop on Intelligent and Evolutionary Systems, AJJWIES 2002*. Canberra: University of NSW.
39. 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
40. 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

41. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. Retrieved from http://arxiv.org/abs/1702.08608

42. 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

43. 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

44. Moser, S. J. (2019). Wandering in Dementia and Trust as an Anticipatory Action. *Medical Anthropology*, *38*(1), 59–70. https://doi.org/10.1080/01459740.2018.1465421