Aplicación de inteligencia artificial y técnicas de aprendizaje automático para la evaluación de la susceptibilidad por movimientos en masa

Juan Pablo Ospina-Gutiérrez* y Edier Aristizábal**

Departamento de Geociencias y Medio Ambiente, Facultad de Minas, Universidad Nacional de Colombia, Av. 80 # 65-223, C.P. 050001, Medellín, Colombia.
**jupospinagu@unal.edu.co, **evaristizabalg@unal.edu.co*

RESUMEN

Los movimientos en masa son uno de los fenómenos de origen natural con mayores pérdidas humanas y económicas alrededor del mundo, por lo que la evaluación de la susceptibilidad y amenaza es una herramienta fundamental para la ordenación de territorios. Existe en la reciente literatura una amplia gama de algoritmos de Inteligencia Artificial con aproximaciones diversas para evaluar y zonificar la susceptibilidad por movimiento en masa. En el presente estudio se implementaron diferentes algoritmos de Aprendizaje Automático para la cuenca de la quebrada La Miel, en los Andes colombianos, con el objetivo de evaluar el desempeño y la capacidad de predicción entre los diferentes modelos. Los resultados obtenidos para este caso de estudio arrojan que los modelos ensamblados tipo *boosting* presentan los mejores valores en términos de desempeño y capacidad de predicción, contrastando con los modelos paramétricos lineales y señalando las limitaciones de éstos en la modelización de problemas complejos, como los movimientos en masa.

Palabras clave: movimientos en masa; susceptibilidad; métodos ensamblados; modelos lineales; aprendizaje automático; Andes colombianos.

ABSTRACT

Landslides are one of the most naturally occurring phenomena with the highest human and economic losses around the world, reason for the susceptibility and hazard assessment is a fundamental tool for land use planning. There is a wide range of Artificial Intelligence algorithms in the recent literature with completely different approaches to establish the relationship between the independent variable (predictors) and the dependent variable (landslide inventory). In the present study, a wide range of algorithms were used for the La Miel creek basin, in the Colombian Andes, and the methodology implemented for this type of data-based modeling is presented in detail and step by step. The results obtained show that the assembled boosting models present the best values in terms of performance and predictability. Contrasting with the linear parametric models, pointing out their limitations in modeling complex problems such as landslides.

Key words: landslide; susceptibility; assembled methods; linear models; machine learning; colombian Andes.

INTRODUCCIÓN

Los movimientos en masa (MenM) son responsables de cerca del 17 % de las muertes ocasionadas por amenazas naturales alrededor del mundo (Lacasse *et al.*, 2010). Estas pérdidas humanas se dan predominantemente en países en vías de desarrollo por la combinación de diferentes factores geológicos y antrópicos que, bajo ciertas condiciones hidrometeorológicas, desencadenan movimientos en masa (Cuanalo *et al.*, 2006; Goetz *et al.*, 2011; Moreno *et al.*, 2006). En contraste, países desarrollados como Estados Unidos y Japón reportan pocas pérdidas humanas, pero altas pérdidas económicas anuales, estimadas entre 1 y 6 mil de millones de dólares (Alimohammadlou *et al.*, 2013).

Entre los países que conforman la región Andina, Colombia es uno de los más afectados por este tipo de amenaza natural, debido a factores como las condiciones tropicales húmedas, derivadas de su localización ecuatorial, su ambiente tectónico de convergencia de tres placas (Nazca, Sur América, Caribe), y su densa ocupación poblacional de áreas susceptibles a la ocurrencia de movimientos en masa (Aristizábal y Sanchez, 2020; Sepúlveda y Petley, 2015).

Según la Base de Datos Internacional de Desastres (EM-DAT), en el periodo transcurrido entre 1901 y 2011, en Colombia ocurrieron 41 desastres por movimiento en masa ocasionando 3,171 víctimas fatales, es decir, 77.34 víctimas por evento, superado únicamente por Perú en donde ocurrieron 318 víctimas por evento (Mergili *et al.*, 2015). Aristizábal y Sánchez (2020), utilizando la base de datos del DesInventar y el Sistema de Información de Movimientos en Masa (SIMMA) del Servicio Geológico Colombiano (SGC), destacan que en el periodo comprendido entre los años 1900 y 2018 han ocurrido al menos 30730 movimientos en masa en Colombia, que han dejado un saldo de 31198 víctimas fatales y pérdidas económicas por USD$ 654 millones.

De manera que la zonificación de la susceptibilidad por movimientos en masa resulta un insumo fundamental para la ordenación y planificación de los territorios. Existen en la literatura múltiples métodos para evaluar la susceptibilidad o amenaza por movimientos en masa, los cuales pueden ser divididos en cuantitativos y cualitativos (Aleotti y Chowdhury, 1999; Carrara *et al.*, 1995; Guzzetti *et al.*, 1999; Van Westen *et al.*, 2006). Los métodos cualitativos utilizan técnicas de cartografía geomorfológica y álgebra de mapas, mediante la asignación de pesos relativos de las variables que influyen en la ocurrencia de movimientos en masa. Es decir, se basan en el conocimiento o criterio de algún experto en el tema y, por lo tanto, son altamente subjetivos (Barredo *et al.*, 2000; Castellanos Abella y Van Westen, 2001;

Ospina-Gutiérrez, J.P., Aristizábal, E., 2021, Aplicación de inteligencia artificial y técnicas de aprendizaje automático para la evaluación de la susceptibilidad por movimientos en masa: Revista Mexicana de Ciencias Geológicas, v. 38, núm. 1, p. 43-54.

DOI: http://dx.doi.org/10.22201/cgeo.20072902e.2021.1.1605

van Westen *et al.*, 2003;). Los métodos cuantitativos pueden ser clasificados en métodos basados en datos o estadísticos, y métodos con base física, o también conocidos como métodos determinísticos o probabilísticos (Palacio *et al.*, 2020). Los métodos con base física, en general, utilizan modelos geotécnicos acoplados con modelos hidrológicos, dando como resultado un factor de seguridad o probabilidades de ocurrencia. Su principal desventaja es que necesitan información geotécnica detallada, por lo que no son recomendables para grandes escalas (Borga *et al.*, 1998; Pourghasemi y Rahmati, 2018; Sorbino *et al.*, 2010; Zieher *et al.*, 2017). Los métodos basados en datos tratan de explicar la relación entre factores de inestabilidad conocidos y la distribución pasada y presente de movimientos en masa, estableciendo relaciones entre variables explicativas (pendiente, geología, aspecto, etc.) y una variable dependiente (ocurrencia de movimientos en masa). Sin embargo, los métodos estadísticos más utilizados en los últimos años hacen asunciones estadísticas sobre la distribución de los datos, lo que los convierte en modelos paramétricos que consideran estrictamente funciones lineales.

Como parte de los métodos estadísticos, en los últimos años se han desarrollado aceleradamente técnicas de Inteligencia Artificial (IA), entre las cuales destaca el Aprendizaje Automático (ML por sus siglas en inglés, *Machine Learning*) que se caracteriza por ser una amplia colección de algoritmos utilizados para crear modelos que aprendan de los datos históricos, con el objetivo de hacer predicciones o conocer las relaciones que pueden existir entre variables de entrada y salida (ocurrencia de un evento); estos modelos tienen varias ventajas, tales como no hacer ninguna asunción estadística de los datos, considerar que no existe linealidad entre las variables y permitir el uso de un amplio tipo de variables (Alpaydin, 2010; Beam y Kohane, 2018; El Naqa y Murphy, 2015; Moreno, 1994). Los algoritmos buscan aprender y extraer conocimiento para formar una función objetivo hipotética (f) que relaciona las variables de entrada o predictoras (x_i) con la variable de salida o variable objetivo (y), como se observa en la ecuación 1.

$$y = f(x_i) \quad (1)$$

Como en muchas otras ramas, estas técnicas han sido implementadas para el desarrollo de modelos de susceptibilidad por movimientos en masa (Pham *et al.*, 2016; Pourghasemi y Rahmati, 2018; Reichenbach *et al.*, 2018). Para este caso, las variables predictoras corresponden a variables del terreno, variables de los materiales de la ladera y variables ambientales.

Una gran cantidad de variables se han utilizado en la literatura (Van Westen *et al.*, 2008). La variable de salida (y), se refiere a la ocurrencia de movimientos en masa identificados, para lo cual se utiliza generalmente un inventario de movimientos en masa (Guzzetti *et al.*, 2012). Para problemas donde la variable de salida, es decir el inventario, no se tenga y sólo se cuente con un conjunto de variables predictoras, se implementan métodos no supervisados, y tienen como objetivo identificar patrones entre los datos (análisis tipo *cluster*), o reducir el número de variables (Análisis de Componentes Principales). Para movimientos en masa, este tipo de métodos de ML no permiten construir modelos de predicción, sin embargo, son muy útiles para seleccionar y evaluar la importancia de las diferentes variables predictoras a utilizar en el modelo (Melchiorre *et al.*, 2008; Sabatakakis *et al.*, 2013). Para problemas donde se cuente tanto con las variables de entrada y salida, se implementan métodos supervisados, que están especializados en predecir nuevos eventos a partir del entrenamiento y aprendizaje que obtuvo el modelo con los datos históricos (Moreno, 1994). Estos métodos se clasifican de acuerdo con la característica de la variable objetivo (y). Los casos en que la variable de salida es de carácter continua, se denominan problemas de regresión y, cuando la variable de salida es de carácter categórica, se denominan problemas de clasificación (Alpaydin, 2010). De acuerdo con el número de posibilidades que pueda tomar la variable objetivo, se denomina clasificación binaria o multiclase. Para el caso de modelos de susceptibilidad por movimiento en masa, donde por definición tienen que predecir la ocurrencia espacial de movimientos en masa, se debe contar entonces con un inventario, el cual es transformado a una variable categórica, con valores de 1 en celdas con presencia de movimientos en masa y 0 con no presencia. Por lo tanto, corresponden a modelos supervisados de clasificación binarios.

En el presente trabajo, se utilizó como caso de estudio la cuenca de la quebrada La Miel, en el municipio de Caldas (Antioquia), para implementar y explorar los diferentes métodos de Aprendizaje Automático para la evaluación de la susceptibilidad por movimiento en masa. Lo anterior con el objetivo de establecer un desarrollo metodológico que considere las ventajas y limitaciones de cada uno de los métodos disponibles. Para evaluar la capacidad de predicción se utilizaron las métricas de validación, matriz de confusión, curva ROC (*Receiver Operating Characteristic*, por sus siglas en inglés), y área bajo la curva AUC (*Area Under Curve*, por sus siglas en inglés).

ZONA DE ESTUDIO

La cuenca de la quebrada La Miel se localiza en la jurisdicción del municipio de Caldas en el departamento de Antioquia. Tiene un área estimada de 22.3 km² (Figura 1) y una temperatura promedio de 19 °C. Se encuentra rodeada del relieve montañoso correspondiente a la Cordillera Central de los Andes con elevaciones sobre el nivel del mar entre 1730 m y 3110 m, es bañada por numerosas fuentes hídricas y vierte sus aguas al Río Medellín (AMVA, 2007).

Según Matula (1981) los horizontes de meteorización de cualquier tipo de roca pueden ser descritos como: I roca fresca, II ligeramente meteorizada, III moderadamente meteorizada, IV altamente meteorizada, V completamente meteorizada y VI suelo residual. En la zona de estudio afloran unidades metamórficas Precámbricas con textura gnéisica desarrollando horizontes de meteorización desde I (roca fresca) hasta VI (suelo residual) y esquistos moscovíticos, cuarzo-sericíticos, grafitosos con intercalaciones de cuarcitas y gneisses los cuales desarrollan horizontes I, II, IV, V. Las unidades cuaternarias corresponden a depósitos aluviales asociados a terrazas del río Medellín con alto grado de redondez y depósitos aluviotorrenciales que se caracterizan por tener una mala selección y una disposición caótica de clastos metamórficos (AMVA, 2006; Mejia, 1984).

METODOLOGÍA

La secuencia metodológica implementada en el presente estudio se resume en los siguientes cinco pasos (Figura 2): (i) Elaboración del inventario de movimientos en masa, (ii) Análisis exploratorio de datos y selección de variables, (iii) Aplicación de algoritmos de ML, (iv) Validación de los modelos y optimización de hiperparámetros (Figuras 3, 4, 5, 6 y 7), y (v) Generación de mapas de susceptibilidad, (Figura 8).

Inventario de movimientos en masa

Para la elaboración del inventario de movimientos en masa de la zona de estudio se realizó una fotointerpretación, utilizando un estereoscopio de espejos con fotografías aéreas de los años 2010-2011 a escala 1:10000. Para la digitalización se utilizó la ortofotografía de la cuenca y se digitalizaron como líneas las coronas de cada uno de los

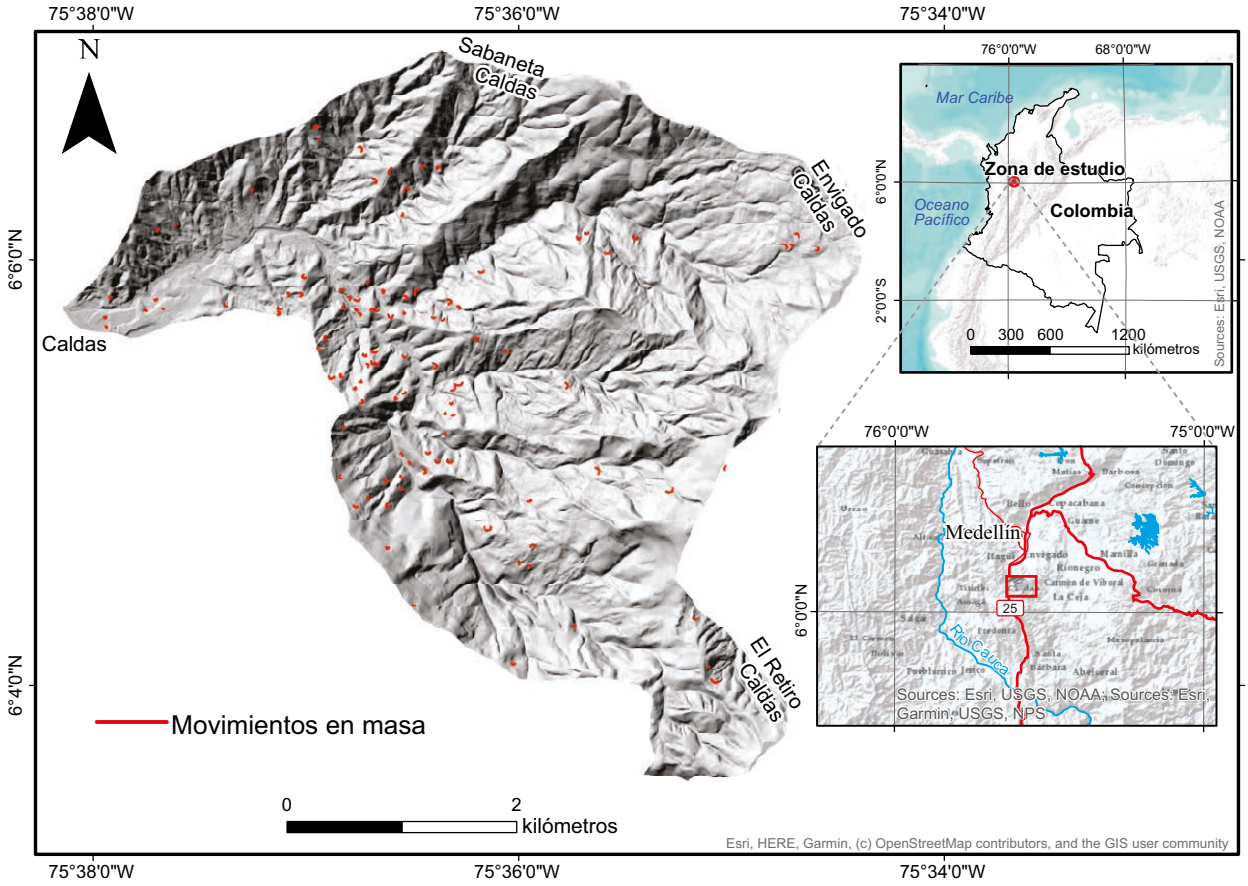

Figura 1. Localización de la zona de estudio en los Andes Colombianos e inventario de movimientos en masa. En color rojo se marcan las coronas de los movimientos en masa inventariados.

movimientos en masa identificados en las fotografías aéreas, considerando que dichas características representan las condiciones previas a la ocurrencia del evento. Se identificaron 105 movimientos en masa tipo planares y superficiales, la mayoría de ellos de tipo complejo, según la clasificación de Cruden y Varnes (1996), ya que presentan una fase posterior a su ocurrencia de tipo flujo. El inventario fue complementado con 32 movimientos en masa históricos consultados en el Sistema de Información de Movimientos en Masa (http://simma.sgc.gov.co/) del Servicio Geológico Colombiano (SGC) (Figura 1).

Variables predictoras

Las variables predictoras fueron inicialmente seleccionadas de acuerdo con Reichenbach *et al.*, (2018). Se construyeron un total de 14 variables predictoras, de tipo continuas y categóricas utilizando el software Arcgis 10.5. Las variables continuas fueron: pendiente, aspecto, rugosidad, curvatura de perfil, curvatura plana, curvatura estándar, elevación, *Stream Power Index* (SPI), *Topographic Wetness Index* (TWI) y acumulación de flujo; dichas variables fueron obtenidas a partir del Modelo Digital de Elevación con resolución espacial de 5 m x 5 m elaborado por el Instituto Geográfico Agustín Codazzi (IGAC) en el proyecto denominado CartoAntioquia. Las variables categóricas utilizadas fueron: geología, elaborada a escala 1:10000 por el estudio de Microzonificación Sísmica del Valle de Aburrá (AMVA, 2006), distancia a fallas, distancia a drenajes, y finalmente el mapa de coberturas del suelo elaborado con las fotografías aéreas de los años 2010-2011 las cuales fueron georreferenciadas y digitalizadas (Tabla 1). Algunas de las variables utilizadas pueden observarse en la Figura 3.

Análisis exploratorio de datos y selección de variables

Uno de los pasos más importantes que permite asegurar resultados óptimos al final del procedimiento, se basa en la revisión, depuración y selección de variables. Para esto, se realizó un análisis exploratorio de los datos, que consistió en la elaboración de histogramas con y sin movimientos en masa para determinar qué variables mostraban diferencias en sus valores ante un movimiento en masa; se elaboraron y analizaron diagramas de dispersión y matriz de correlación con el fin de observar qué variables se encontraban correlacionadas para evitar incluir ruido en los modelos y, finalmente, se usó la técnica no supervisada de Análisis de Componentes Principales, que reduce la dimensionalidad de un problema (cantidad de variables) seleccionando aquellas variables que recojan la mayor parte de variabilidad de los datos, perdiendo la menor cantidad de información posible.

Algoritmos de aprendizaje automático

Para la aplicación de los modelos, se dividió de manera aleatoria tanto a las variables predictoras (887282 pixeles) como al inventario de movimientos en masa (1453 píxeles), según la regla de Pareto, en 80 % de entrenamiento y 20 % de validación. Es importante notar que la ocurrencia de movimientos en masa representa sólo el 0.16 % de todos los píxeles de la cuenca, por lo que se considera un problema desbalanceado en ML. Si se entrena un modelo en el cual la clase que se quiere predecir es minoritaria, el modelo no aprenderá adecuadamente y las predicciones serán sólo de la clase mayoritaria, arrojando altos valores en las métricas de evaluación (Guo y Viktor, 2004). Para afrontar este problema, se utilizaron dos técnicas de remuestreo: (i) submuestreo,

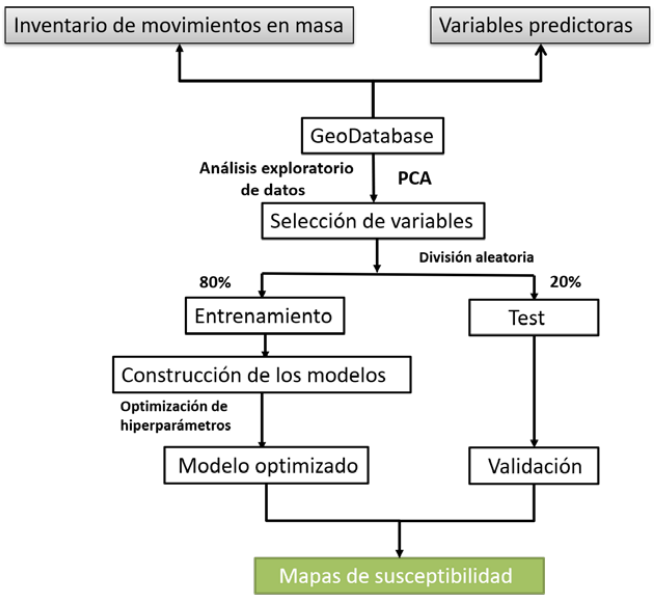

Figura 2. Flujo de trabajo utilizado. PCA: Análisis de componentes principales.

utilizando la función RUS (por sus siglas en inglés *Random Under Sampler*), y (ii) sobre-muestreo, con la función SMOTE (por sus siglas en inglés *Synthetic Minority Over-sampling Technique*) las cuales están disponibles en la librería de *Scikit-learn* de Python 3.7.

Con las celdas de entrenamiento se implementaron 11 modelos de ML utilizando la librería *Scikit-learn* de Python 3.7. Cada uno de los modelos representa una estrategia diferente de aprendizaje y pueden agruparse en: (1) Métodos paramétricos, divididos en (a) métodos no supervisados lineales: Análisis de componentes principales (PCA) (Abdi y Williams**,** 2010; Smith, 2002; Villardón, 2002), y (b) métodos supervisados lineales: Regresión Logística (RL) (López y Facheli, 2015) y Análisis Discriminante Lineal (LDA) (Ramos *et al.*, 2016); (2) Métodos no paramétricos no lineales: dividido en K Vecinos más Cercanos (KNN) (Marjanovic *et al*., 2009), Árboles de Decisión (DT) (Friedl y Brodley, 1997; Myles *et al.*, 2004; Raileanu y Stoffel, 2004), Máquinas de Vectores de Soporte (SVM) (Carmona, 2013) y Redes Neuronales Artificiales tipo Perceptrón multicapas (ANN) (Basogain, 1998; Gardner y Dorling, 1998; Noriega, 2005; Matich, 2001); y (3) Métodos tipo ensamble *Bagging*: Bosques Aleatorios (RF) (Gislason *et al.,* 2006), y tipo *Boosting*: Adaboost (ADB) (Dietterich, 2000), Gradiente Estocástico (SGB) (Brown y Mues, 2012; Friedman, 2002), y el denominado XGboost (XGB) (Chen y Guestrin, 2016).

Los modelos paramétricos comúnmente utilizados para la evaluación de la susceptibilidad por movimiento en masa simplifican el aprendizaje a partir de los datos de entrenamiento, ya que utilizan una función lineal que generalmente corresponde a una combinación de las variables predictoras (Alpaydin, 2010). Estos modelos generalmente no presentan problemas de varianza, sin embargo, en muchos casos, el problema a modelar no se ajusta a una función lineal, por lo que pueden presentar problemas de ajuste o sesgo importantes (Kuhn y Johnson, 2013). Es por esta razón que los modelos de Aprendizaje Automático presentan nuevos algoritmos que utilizan modelos no paramétricos, ya que no asumen ninguna forma previa de la función objetivo, permitiendo aprender en teoría a partir de los datos de entrenamiento cualquier forma de la función. Generalmente, este tipo de modelos no presenta problemas de ajuste o sesgo, pero en su defecto tienden a sobre ajustarse a los datos, generando problemas de varianza ante nuevas observaciones (Mehta *et al.,* 2019). Esto significa una importante ventaja, pero al mismo tiempo exige construir un modelo con un balance entre el ajuste y la varianza (Alpaydin, 2010; Belkin *et al.,* 2019; Srinivasan y Fisher, 1995; Geman *et al.,* 1992).

Una de las técnicas más utilizadas actualmente en ML para mejorar la capacidad de generalización y predicción del modelo, con excelentes resultados, se refiere al ensamblaje de métodos con estimadores base. Las dos técnicas de ensamblaje más utilizadas son denominadas *bagging* y *boosting*. La diferencia entre ellas consiste en que la técnica *bagging* agrega una serie de estimadores base independientes y en forma simultánea, mientras que la *boosting* agrega los estimadores simples de forma secuencial y dependiente. De manera que, las técnicas tipo *bagging* se utilizan para estimadores robustos e inestables que tienden a sobre ajustarse y tener problema de varianza, mientras que las técnicas *boosting* se utilizan, por el contrario, con estimadores débiles que presentan problemas de ajuste (Dietterich, 2000; Mehta *et al.,* 2019).

Validación de los modelos y optimización de hiperparámetros

En los modelos de ML existen parámetros e hiperparámetros. Los parámetros dependen de los datos y son utilizados para realizar predicciones por lo que su valor es estimado durante el entrenamiento, por ejemplo, los coeficientes de las variables en regresión logística. Por el contrario, los hiperparámetros son establecidos por el usuario, es decir, no son estimados de los datos (Kuhn y Johnson, 2013). Habitualmente, se recurre a la experiencia y al ensayo y error para establecerlos, aunque también existen métodos más refinados como la optimización, la cual consiste en establecer una grilla con diferentes valores y buscar la mejor combinación posible a través de validación cruzada. Para esto existen funciones como *Grid Search* y *Random Search* (Bergstra y Bengio, 2012).

Para la fase de entrenamiento se implementaron cada uno de los modelos con el 80 % de la geodatabase, realizando una optimización de hiperparámetros con el algoritmo de *Random Search* incluido en la librería de *Scikit-learn*, y se hizo una validación cruzada en la cual se generan modelos seleccionando aleatoriamente hiperparámetros, subdividiendo las observaciones en un número *k*, donde *k*-1 subdivisiones son utilizadas para entrenar el modelo, y la subdivisión restante es utilizada para validar los resultados. Este procedimiento se realiza *k* veces, lo cual permite que todos los datos de entrenamiento sean utilizados tanto para entrenar como para validar el modelo.

Para la evaluación del modelo se utilizó la Matriz de Confusión y las métricas derivadas: Curva ROC, área bajo la curva ROC (AUC) y el *Recall* (Fawcett, 2006). El desempeño del modelo se refiere a los valores obtenidos de las métricas con los datos de entrenamiento, y señalan la respuesta del modelo a los datos de entrada. En tanto, la capacidad de predicción del modelo corresponde a los valores obtenidos utilizando los datos de validación no utilizados en el entrenamiento (20 %). Tanto el desempeño como la predicción de cada uno de los modelos fue evaluada.

Generación de mapas de susceptibilidad

Una vez que cada uno de los modelos fue validado y que se seleccionaron sus mejores hiperparámetros, se procedió a predecir la totalidad de la cuenca (887282 píxeles). Un paso importante para obtener el mapa final de susceptibilidad es transformar el mapa de valores continuos de probabilidad, obtenido por el modelo aplicado, a un mapa tipo semáforo, representando la susceptibilidad en categorías. Para este estudio se seleccionó el mejor umbral en los modelos como la mínima distancia a la clasificación perfecta en la curva ROC, que corresponde al punto (0.1). Este umbral optimiza los TPR (*True Positive Rate*, por sus siglas en inglés) de acuerdo con los FPR *(False Positive Rate*), es decir, encuentra un balance del modelo entre los verdaderos positivos contra los falsos positivos (Figura 8). Sin embargo, para clasificar la

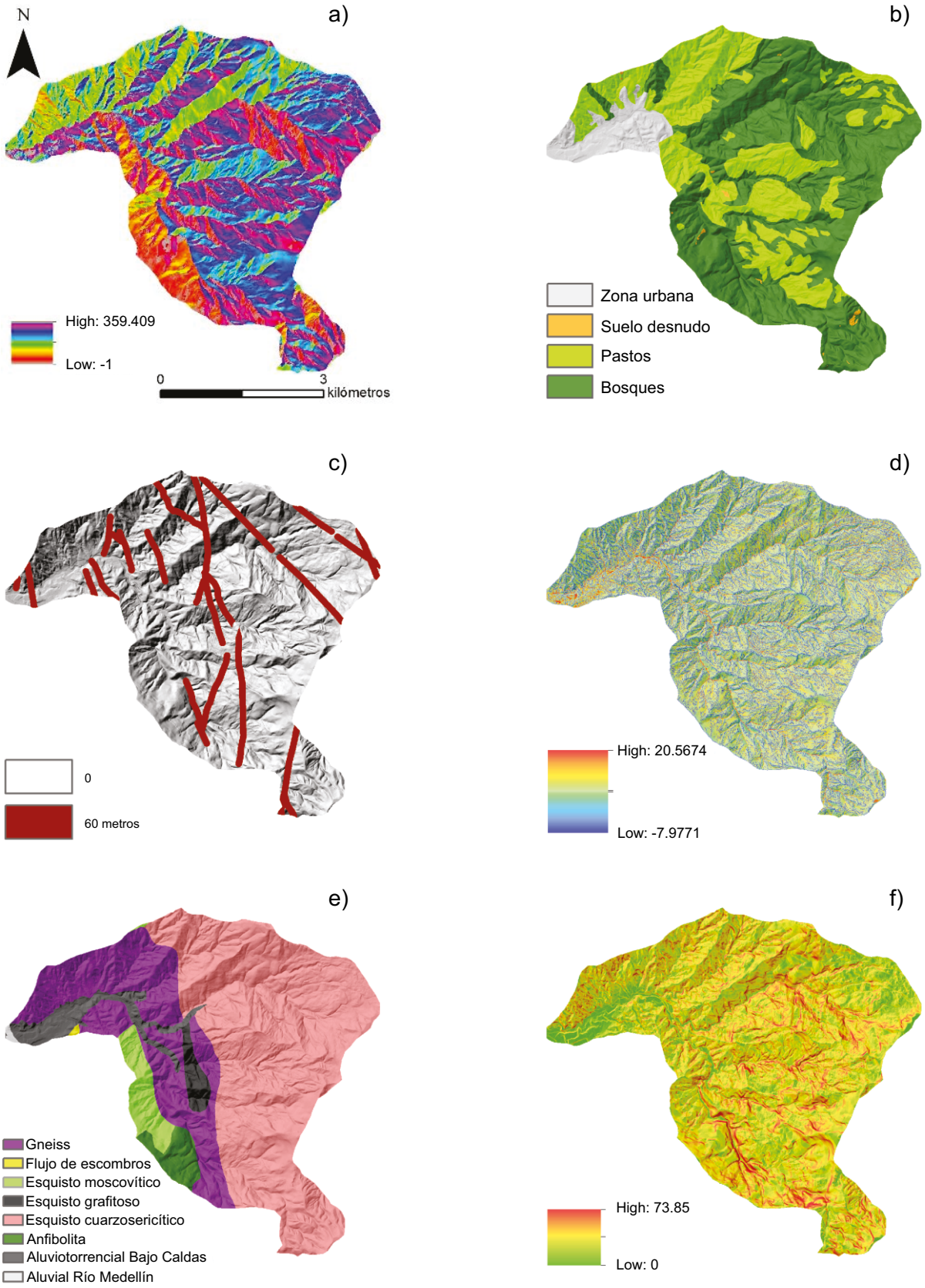

Figura 3. Variables predictoras: (a) Aspecto, (b) Coberturas, (c) Buffer lineamientos, (d) TWI, (e) Geología, (f) Pendiente.

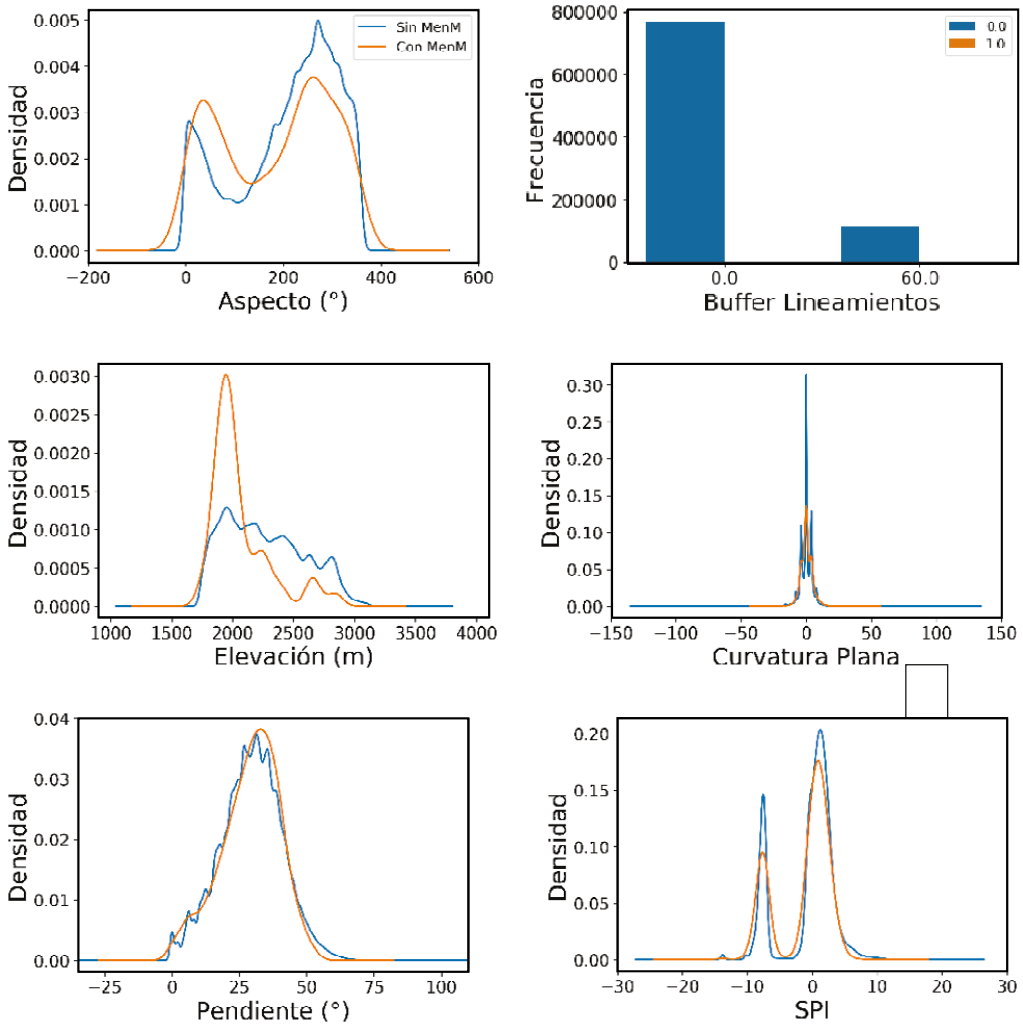

Figura 4. Histogramas diferenciados por celdas con y sin movimientos en masa. MenM: Movimiento de masa; SPI: índice de erosión.

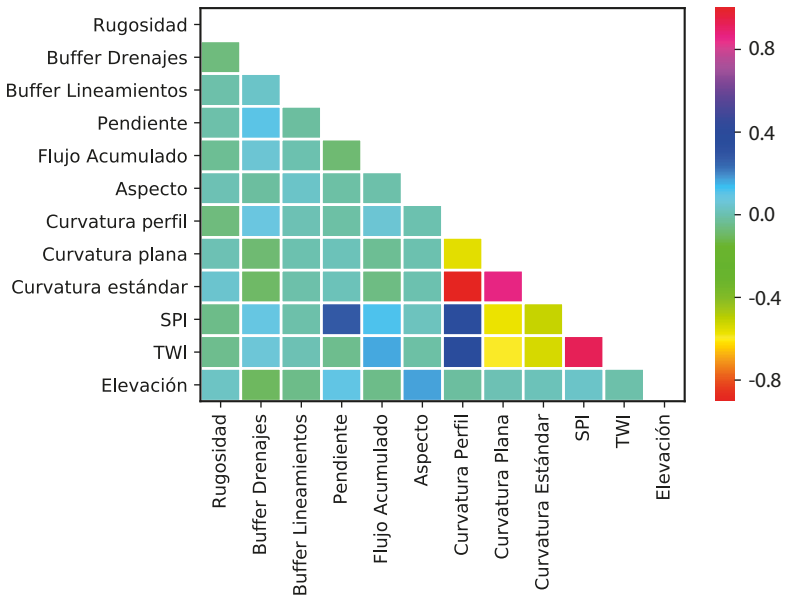

Figura 5. Matriz de correlación de Pearson entre todas las variables predictoras. SPI: índice de erosión; TWI: índice de humedad topográfico.

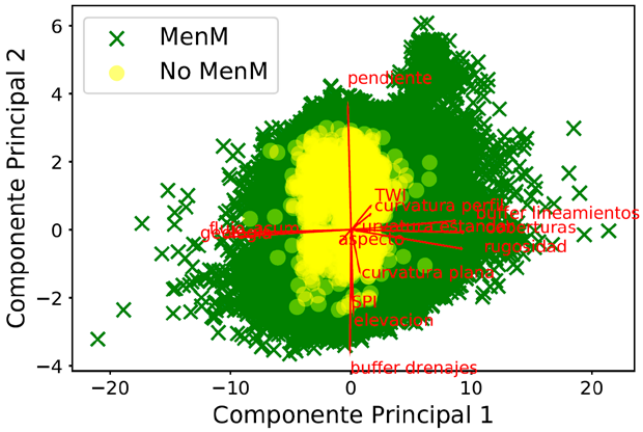

Figura 6. *Biplot* del Análisis de Componentes Principales para los componentes principales (CP) 1 y 2 en donde se observa que las variables que más aportan a la varianza del problema son: buffer lineamientos, coberturas, rugosidad, curvatura estándar, geología, flujo acumulado, curvatura perfil en el CP1 y pendiente, elevación, SPI, buffer drenajes en el CP2. MenM: movimientos en masa.

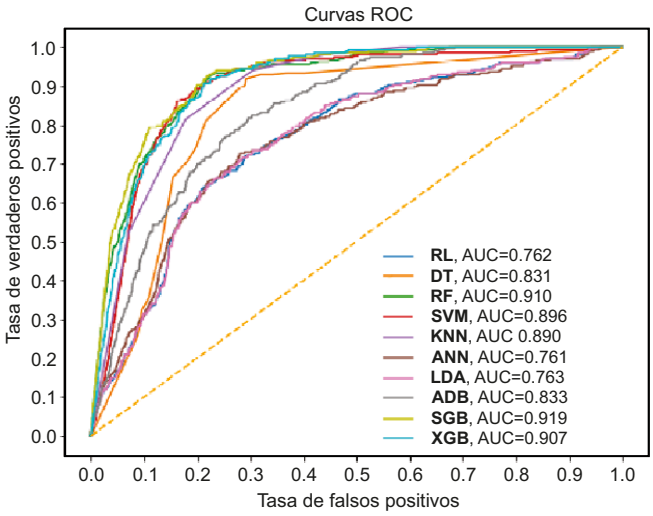

Figura 7. Curvas ROC y AUC de cada uno de los modelos. RL=Regresión logística, DT=Árboles de decisión, RF=Bosques Aleatorios, SVM=Máquinas de Vector de Soporte, KNN=K vecinos más cercanos, ANN=Redes Neuronales Artificiales tipo Perceptrón multicapas, LDA=Análisis Discriminante Lineal, ADB=Adaboost, SGB=Gradiente Estocástico, XGB=XGboost.

susceptibilidad media y baja en el modelo con los mejores resultados se utilizó como criterio la probabilidad espacial del 5 % para la susceptibilidad baja. Es decir, las áreas de susceptibilidad baja presentan una probabilidad espacial de ocurrencia de movimientos en masa del 5 %, la de amenaza media del 13 % y la amenaza alta del 82 %.

RESULTADOS

Para la selección de variables se realizaron histogramas de cada una de las variables diferenciando las celdas con deslizamientos y las celdas con no deslizamientos (Figura 4). Las variables elevación, TWI, SPI, curvaturas (plana, perfil, estándar), aspecto, flujo acumulado, buffer drenajes y geología exhiben un comportamiento distinto cuando ocurre un movimiento en masa, por lo que se consideraron como buenas variables predictoras; caso contrario ocurre con la rugosidad, pendiente y buffer lineamientos ya que no se diferenciaron en términos de movimientos en masa, por lo que no son variables predictoras adecuadas. Se analizó el nivel de correlación mediante el coeficiente de correlación de Pearson.

La Figura 5 presenta la matriz de correlación de Pearson, entre todas las variables. En general, se observan valores de correlación bajos y positivos. Sólo las variables curvatura plana y curvatura estándar presentan una correlación negativa media con las variables SPI y TWI; y la variable curvatura perfil una correlación positiva media con curvatura estándar y curvatura plana. En cuanto al análisis de Componentes Principales (Figura 6), los seis primeros componentes explican aproximadamente el 75 % de la varianza del problema donde los que más aportan a la varianza son: curvatura estándar, geología, flujo acumulado, curvatura perfil, pendiente, elevación, SPI, y buffer drenajes.

Con base en la integración de los análisis anteriormente explicados, se seleccionaron las siguientes variables para construir el modelo: Pendiente, Geología, Curvatura de perfil, TWI, Aspecto, Buffer drenajes, Buffer lineamientos y Elevación.

En la Figura 7 puede observarse la curva ROC y AUC de cada uno de los modelos, donde resaltan SGB (0.919), XGB (0.907), RF (0.91), SVM (0.896), KNN (0.890), ADB (0.833) y DT (0.831) con los valores más altos usando únicamente el submuestreo de RUS. No fue posible ejecutar estos modelos usando el remuestreo SMOTE, debido a la gran cantidad de datos y la poca capacidad computacional disponible. En la Tabla 2 puede observarse el *Recall* con el AUC de cada modelo, tanto para el desempeño como para la capacidad de predicción. Los modelos con menor capacidad de predicción presentan valores menores de AUC y *Recall*. En la Figura 8 se presentan los mapas de susceptibilidad por movimientos en masa finales considerando las mejores variables predictoras, y con los hiperparámetros optimizados.

DISCUSIÓN

De acuerdo con los resultados, como parte fundamental para obtener un buen modelo se encuentra la selección de las variables predictoras adecuadas. Una buena variable predictora exhibe un comportamiento estadístico diferenciado entre celdas con y sin

Tabla 1. Resumen de variables predictoras.

Variable	Mín.	Máx.	Promed.	Escala	Fuente
Continuas					
Pendiente	0	73.8	29.4	Ráster con resolución espacial 5 m x 5 m	IGAC, 2012
Aspecto	-1	359	208		
Rugosidad	0	0.98	0.5		
Curvatura de perfil	-85	116	0.09		
Curvatura plana	-67	66	0.01		
Curvatura estándar	-144	152	0		
Elevación	1730	3110	2263		
Stream Power Index (SPI)	-13	13	-0.54		
Topographic Wetness Index (TWI)	-7.9	20.5	0.84		
Flujo acumulado	0	857764	1006		
Categóricas					
Geología				1:10000	IGAC, 2012
Coberturas					
Distancia a fallas					
Distancia a drenajes					

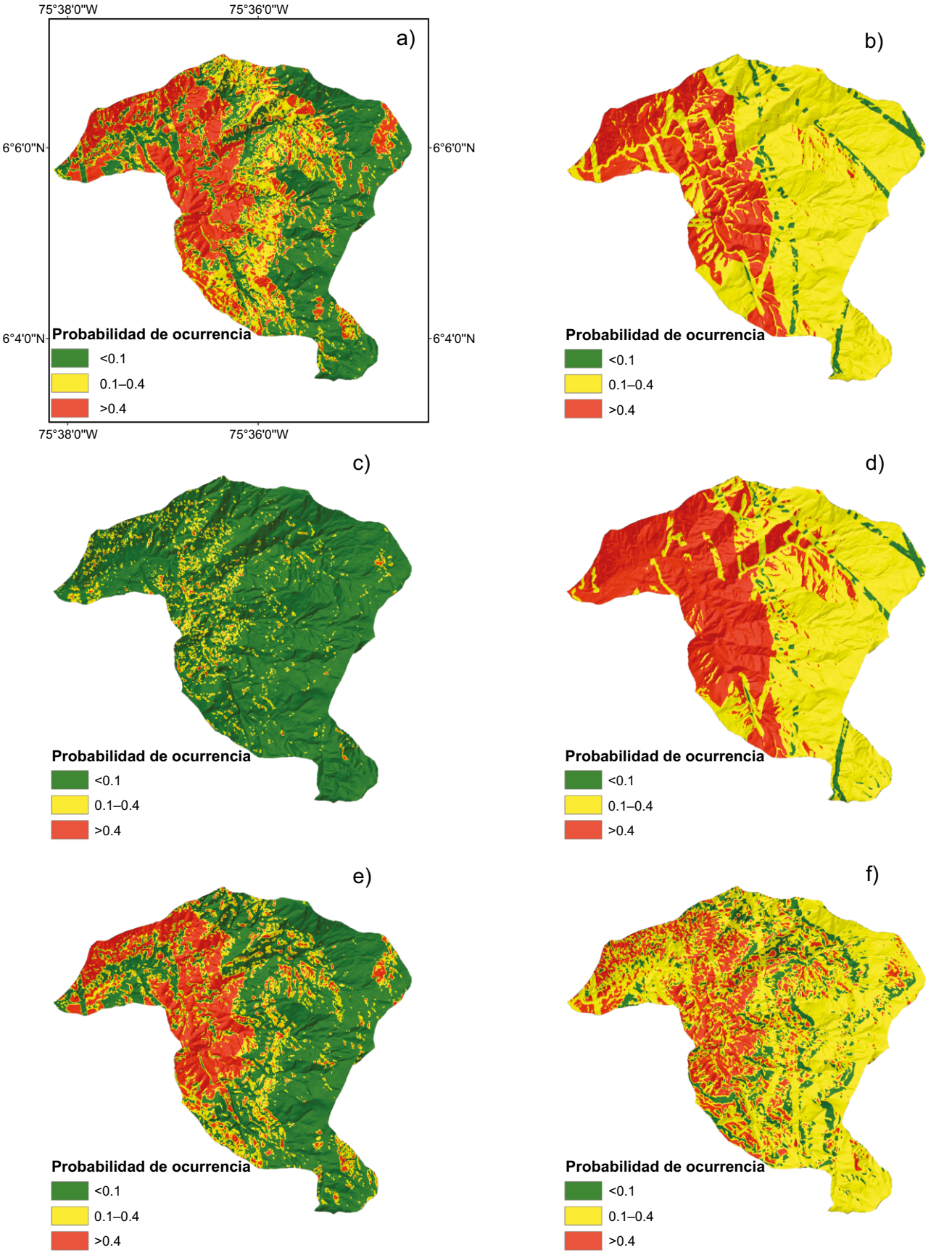

Figura 8. Mapas de susceptibilidad por movimientos en masa: (a) DT, (b) ANN, (c) KNN, (d) RL, (e) SGB, (f) SVM.

Tabla 2. Métricas de validación de los modelos.

Algoritmo	*Recall*		AUC	
	Desempeño	Predicción	Desempeño	Predicción
SGB	1	0.84	1	0.919
RF	1	0.84	1	0.91
SVM	0.99	0.84	0.99	0.896
XGB	1	0.83	1	0.907
DT	0.95	0.81	0.96	0.831
ADB	0.80	0.76	0.84	0.833
LDA	0.72	0.73	0.76	0.763
ANN	0.53	0.72	0.76	0.761
RL	0.72	0.71	0.76	0.762
KNN	0.96	0.52	0.96	0.89

presencia del evento (Aristizábal *et al.,* 2019). En caso contrario dicha variable no discrimina entre celdas estables e inestables. Como se mencionaba, también es importante que la variable correlacione o represente la variable de salida, pero que no correlacione con otras variables predictoras. Por lo que, con base en la integración del análisis exploratorio de datos, compuesto por histogramas, gráficos de dispersión, matriz de correlación y análisis de componentes principales, es posible seleccionar las mejores variables predictoras con la información disponible. Cabe destacar que dicho análisis depende de las condiciones locales de la zona de estudio y que cada zona es diferente, en consecuencia, no se obtienen siempre las mismas variables. Por lo que, inicialmente, se puede partir de un rango amplio de variables, basado en el estado del arte y conocimiento de la zona de estudio, para luego reducir dicho rango de variables a aquellas que discriminan mejor y en específico para la zona de estudio.

En cuanto a los algoritmos de ML utilizados, de acuerdo con los resultados obtenidos para el caso de la quebrada La Miel en el municipio de Caldas (Antioquia), los modelos con el mejor desempeño y capacidad de predicción son SGB, RF y XGB. Lo que señala el excelente desempeño de los modelos ensamblados, especialmente para los datos de entrenamiento, mientras que para los datos de validación su capacidad se reduce de forma significativa (0.83–0.84), señalando problemas de varianza y sobreajuste del modelo. La curva de aprendizaje indica que los modelos aprenden rápidamente, tendiendo la curva de validación a incrementar sustancialmente a medida que aumentan las observaciones; al final, la curva tiende a ser asintótica señalando que el modelo está llegando a su máxima capacidad de aprendizaje, pero conservando una diferencia importante entre la curva de entrenamiento y la curva de validación, esta diferencia corresponde a la varianza que conserva el modelo (Figura 9e, 9f).

También es importante considerar que este tipo de algoritmos tienen una amplia variedad de hiperparámetros para optimizar, lo cual es una ventaja en muchos casos, pero que se convierte generalmente en una tarea ardua y que exige una gran capacidad de cómputo.

Los modelos no paramétricos y no lineales, presentan diferentes rendimientos. El mejor modelo es SVM, que a pesar de ser un modelo simple, implementa las funciones kernel, que le permiten ajustarse a espacios de mayores dimensiones y, por lo tanto, de mayor complejidad; sin embargo, el costo computacional de ejecutarlo es alto, por lo que se limita a bases de datos pequeñas, e incluso para nuestro caso no pudo implementarse con el remuestreo de SMOTE. SVM presentó valores muy altos de *Recall* y AUC con los datos de entrenamiento, y su capacidad de predicción se redujo con los datos de validación a 0.84 y un AUC de 0.896, lo que indica problemas de varianza y sobre ajuste. En cuanto a ANN, requiere una ardua optimización sobre la arquitectura y los hiperparámetros que controlan el aprendizaje de la red. La curva de aprendizaje indica que la arquitectura final es altamente inestable y no representa adecuadamente el problema (Figura 9d). KNN presenta muy buenos resultados para los datos de entrenamiento, en términos de *Recall*, conservando valores altos de AUC para los datos de validación (Figura 9c).

En cuanto a los métodos paramétricos lineales RL y LDA, presentan un rendimiento moderado (*Recall*: 0.71–0.73), debido posiblemente a que son modelos que no logran capturar la complejidad de las relaciones entre las variables y la ocurrencia del evento. En cuanto a las curvas de aprendizaje presentan un comportamiento similar, el algoritmo aprende con los datos, pero a una velocidad baja, lo cual no permite obtener valores altos de desempeño y capacidad de predicción (Figura 9b).

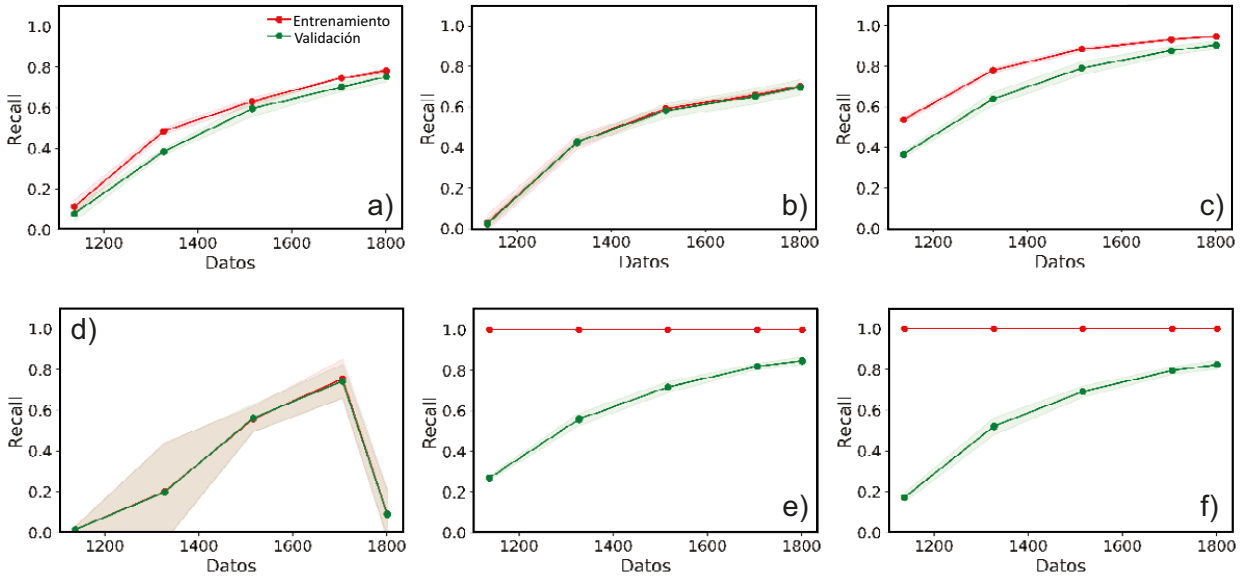

Figura 9. Curvas de aprendizaje: (a) ADB, (b) LDA, (c) KNN, (d) ANN, (e) SGB, (f) RF.

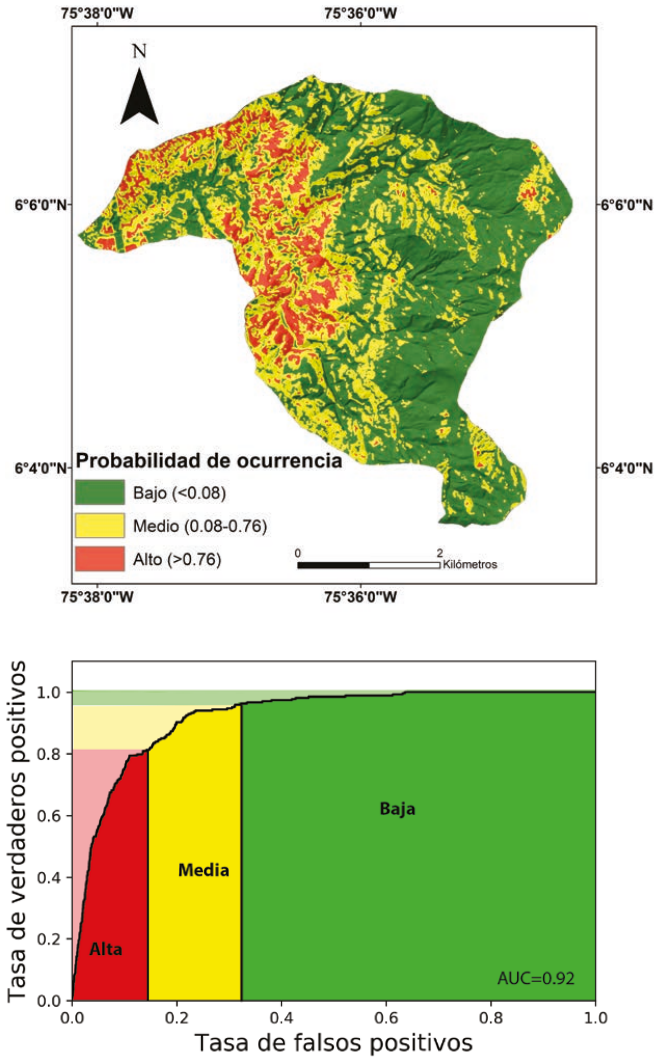

Figura 10. Mapa de Susceptibilidad por movimientos en masa (a) para la cuenca de la quebrada La miel, y curva ROC (b) con los datos de validación.

Es importante resaltar que la selección de métricas adecuadas para la validación de modelos implementados sobre problemas desbalanceados es fundamental, ya que algunas pueden dar una falsa expectativa sobre el rendimiento de cualquier modelo. Por consiguiente, es importante utilizar diferentes métricas para validar los modelos y, en especial, aquellas que hagan énfasis en la correcta clasificación de la clase de interés. En casos donde el problema físico a modelar es desbalanceado por naturaleza, la función *Recall* funciona adecuadamente, ya que se centra en la ocurrencia del evento.

Finalmente, en la Figura 10 se presenta el mapa de susceptibilidad por movimientos en masa para la quebrada La Miel del municipio de Caldas utilizando el algoritmo SGB y el umbral óptimo. Los resultados señalan que el 84 % de las celdas de la cuenca corresponden a verdaderos positivos y el 16 % a falsos positivos, los falsos negativos representan menos del 1 %. Los resultados señalan una alta eficiencia del modelo, ya que tan sólo cerca del 16 % del área de la cuenca contiene el 82 % de los movimientos en masa históricos. Lo que significa que para identificar adecuadamente los movimientos en masa o zonas inestables no se requiere clasificar como de susceptibilidad alta a amplias zonas de la cuenca.

CONCLUSIONES

De acuerdo con los resultados obtenidos en el presente estudio, las nuevas herramientas proporcionadas por la Inteligencia Artificial y Aprendizaje Automático se destacan como una excelente alternativa que proporcionan modelos que se pueden ajustar adecuadamente a la realidad del problema enfrentado. Sin embargo, al igual que para las técnicas clásicas de la evaluación de la susceptibilidad por movimientos en masa como métodos heurísticos o con base física, se requiere tener un buen inventario de movimiento en masa, y un detallado levantamiento de las variables predictoras. Estas condiciones básicas permitirán obtener y ajustar modelos útiles para la toma de decisiones y entendimiento del fenómeno.

En cuanto al desempeño de los algoritmos disponibles de Aprendizaje Automático, los modelos no paramétricos arrojan los mejores resultados, basados en su capacidad de adaptarse o aprender de los datos. Para lo cual requieren una base de datos extensa. Dentro de estos modelos, los algoritmos ensamblados mostraron los mejores valores en términos de desempeño y capacidad de predicción. En este sentido es importante destacar que los modelos paramétricos lineales, como RL y LDA que son los modelos más ampliamente utilizados en la evaluación de la susceptibilidad por movimiento en masa, arrojan los resultados más pobres. Esto puede deberse a la restricción de linealidad en los modelos, lo cual seguramente no se cumple para modelar fenómenos altamente complejos y de múltiples factores causales, como los movimientos en masa.

Finalmente, los resultados señalan que la cuenca baja de la quebrada La Miel presenta los mayores índices de susceptibilidad por movimientos en masa, asociados a laderas de moderadas y fuertes pendientes, cubiertas por pastos y conformadas por rocas metamórficas tipos gneis, en las cuales el macizo rocoso presenta estructuras tipo esquistosidad que favorece la ocurrencia de movimientos en masa tipo planar.

REFERENCIAS

Abdi, H., Williams, L.J., 2010, Principal component análisis: Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

Aleotti, P., Chowdhury, R., 1999, Landslide hazard assessment: summary review and new perspectives: Bulletin of Engineering Geology and the Environment, 58(1), 21-44.

Alimohammadlou, Y., Naja, A., Yalcin, A., 2013, Landslide process and impacts : A proposed classification method: CATENA, 104, 219-232.

Alpaydin, E., 2010, Introduction to Machine Learning: Cambridge, MIT press, 517 pp.

Área Metropolitana del Valle de Aburrá (AMVA), 2006, Microzonificación sísmica detallada de los municipios De Barbosa, Girardota, Copacabana, Sabaneta, La Estrella, Caldas Y Envigado: Medellín, Colombia, reporte técnico, 679 pp.

Área Metropolitana del Valle de Aburrá (AMVA), 2007, Plan de Ordenamiento y Manejo de la Cuenca del Río Aburrá: Medellín, Colombia, reporte técnico, 171 pp.

Aristizábal, E., Sanchez, O., 2020, Spatial and temporal patterns and the socioeconomic impacts of landslides in the tropical and mountainous Colombian Andes: Disasters, 44(3), 596-618. doi: 10.1111/disa.12391.

Aristizábal, E., López, S., Sánchez, O., Vásquez, M., Rincón, F., Ruiz-Vásquez, D., Valencia, J. S., 2019, Evaluación de la amenaza por movimientos en masa detonados por lluvias para una región de los Andes colombianos estimando la probabilidad espacial, temporal, y magnitud: Boletín de Geología, 41(3), 85-105.

Barredo, J., Benavides, A., Hervás, J., van Westen, C.J., 2000, Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island, Spain: International Journal of Applied Earth Observation and Geoinformation, 2(1), 9-23.

Basogain, X., 1998, Redes Neuronales Artificiales y sus aplicaciones: Bilbao, España, Publicaciones de la Escuela de Ingenieros, 76 pp.

Beam, A.L., Kohane, I.S., 2018, Big Data and Machine Learning in Health Care: Jama, 319(13), 1317-1318.

Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019, Reconciling modern machine-learning practice and the classical bias – variance trade-off: Proceedings of the National Academy of Sciences, 116(32), 15849-15854.

Bergstra, J., Bengio, Y., 2012, Random search for hyper-parameter optimization: The Journal of Machine Learning Research, 13(1), 281-305.

Borga, M., Dalla Fontana, G., Da Ros, D., Marchi, L, 1998, Shallow landslide hazard assessment using a physically based model and digital elevation data: Environmental geology, 35(2-3), 81-88.

Brown, I., Mues, C., 2012, An experimental comparison of classification algorithms for imbalanced credit scoring data sets: Expert Systems with Applications, 39(3), 3446-3453.

Carmona, E., 2013, Tutorial sobre Máquinas de Vectores Soporte (SVM): Madrid, España, Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial, reporte técnico, 27 pp.

Carrara, A., Cardinali, M., Guzzetti, F., Reichenbach, P., 1995, Gis technology in mapping landslide Hazard: Geographical Information Systems in Assessing Natural Hazards, 5, 135-175.

Castellanos Abella, E.A., van Westen, C.J., 2001, Landslide hazard assessment using the heuristic model (resumen), *en* GEOMIN: Geología y Minería 2001: memorias trabajos y resúmenes: La Habana, Cuba, GECAMIN, 10-20.

Chen, T., Guestrin, C., 2016, XGBoost: A scalable tree boosting system, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA: New York, Association for Computing Machinery, 785-794. doi: 10.1145/2939672.2939785.

Cruden, D.M., Varnes, D.J, 1996, Landslides: investigation and mitigation: Transportation Research Board special report, 247, 39 pp.

Dietterich, T. G., 2000, An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging , Boosting , and Randomization: Machine Learning, 40, 139-157.

El Naqa, I., Murphy, M.J., 2015, What Is Machine Learning? *en* El Naqa, Issam, Li, Ruijiang, Murphy, Martin (eds.) Machine Learning in Radiation Oncology: Theory and Applications, Switzerland, Springer International Publishing, 3-11.

Fawcett, T., 2006, An introduction to ROC análisis*:* Pattern recognition letters, 27(8), 861-874.

Friedl, M., Brodley, C., 1997, Decision Tree Classification of Land Cover from Remotely Sensed Data: Remote Sensing of Environment, 61(3), 399-409.

Friedman, J.H., 2002, Stochastic gradient boosting*:* Computational Statistics and Data Analysis, 38(4), 367-378.

Gardner, M.W., Dorling, S.R., 1998, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences: Atmospheric Environment, 32(14-15), 2627-2636.

Geman, S., Bienenstock, E., Doursar, R., 1992, Neural Networks and the Bias/Variance Dilemma: Neural Computation, 4(1), 1-58.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006, Random Forests for land cover classification: Patter Recognition Letters, 27, 294-300.

Goetz, J.N., Guthrie, R.H., Brenning, A., 2011, Integrating physical and empirical landslide susceptibility models using generalized additive models: Geomorphology, 129(3-4), 376-386.

Guo, H., Viktor, H.L., 2004, Learning from imbalanced data sets with boosting and data generation: the databoost-im approach: ACM Sigkdd Explorations Newsletter, 6(1), 30-39.

Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P., 1999, Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy: Geomorphology, 31(1-4), 181-216.

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K., 2012, Earth-Science Reviews Landslide inventory maps : New tools for an old problema: Earth Science Reviews, 112(1-2), 42-66.

Instituto Geográfico Agustín Codazzi (IGAC), 2012, Toma de fotografías aéreas y restitución cartográfica del departamento de Antioquia a escalas 1:10,000, 1:5,000 y 1:2,000 contratado por el Área Metropolitana del Valle de Aburrá, ISAGEN, Empresas Públicas de Medellín, Municipio de Medellín, Gobernación de Antioquia y el IDEA.

Kuhn, M., Johnson, K., 2013, Applied Predictive Modeling: New York, Springer, 600 pp.

Lacasse, S., Nadim, F., Kalsnes, B., 2010, Living with landslide risk: Geotechnical Engineering Journal of the SEAGS & AGSSEA 41(4).

López, P., Fachelli, S., 2015, Metodología de la investigación social cuantitativa: Barcelona, España, Universidad Autónoma de Barcelona, 56 pp.

Marjanovic, M., Bajat, B., Kovacevic, M., 2009, Landslide susceptibility assessment with machine learning algorithms, *en* 2009 International Conference on Intelligent Networking and Collaborative Systems, Barcelona, España: Los Alamitos, California, IEE, 273-278.

Matula, M., 1981, Rock and soil description and classification for engineering geological mapping report by the IAEG Commission on Engineering Geological Mapping: Bulletin of the International Association of Engineering Geology-Bulletin de l'Association Internationale de Géologie de l'Ingénieur, 24(1), 235-274.

Mehta, P., Bukov, M., Wang, C.H., Day, A. G., Richardson, C., Fisher, C.K., Schwab, D.J., 2019, A high-bias, low-variance introduction to machine learning for physicists: Physics reports, 810, 1-124.

Mejía, N., 1984, Geología y geoquímica de las planchas 130 (Santafé de Antioquia) y 146 (Medellín Occidental), escala 1: 100.000: Medellín, Colombia, Instituto Colombiano de Geología y Minería (INGEOMINAS), memoria explicativa, 397 pp.

Melchiorre, C., Matteucci, M., Azzoni, A., Zanchi, A., 2008, Artificial neural networks and cluster analysis in landslide susceptibility zonation: Geomorphology, 94, 379-400.

Mergili, M., Marchant, C., Moreiras, S.M., 2015, Causas, características e impacto de los procesos de remoción en masa, en áreas contrastantes de la región Andina: Cuadernos de Geografía: Revista Colombiana de Geografía, 24(2), 113-131.

Moreno, A., 1994. Aprendizaje automático: Barcelona, España, Edicions UPC, 342 pp.

Moreno, H., Vélez Ortiz, M., Montoya, J., Rhenals Monterrosa, R., 2006, La lluvia y los deslizamientos de tierra en Antioquia: Análisis de su ocurrencia en las escalas interanual, intraanual y diaria: Revista EIA, 5, 59-69.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004, An introduction to decision tree modeling: Journal of Chemometrics, A Journal of the Chemometrics Society, 18(6), 275-285.

Noriega, L., 2005, Multilayer Perceptron Tutorial: Stafford, Reino Unido, Staffordshire University, School of Computing, reporte técnico, 12 pp.

Palacio, J., Mergili, M., Aristizábal, E, 2020, Probabilistic landslide susceptibility analysis in tropical mountainous terrain using the physically based r. slope. Stability model: Natural Hazards & Earth System Sciences, 20(3), 815-829. doi: 10.5194/nhess-20-815-2020.

Pham, B.T., Pradhan, B., Bui, D.T., Prakash, I., Dholakia, M.B., 2016, Environmental Modelling y Software A comparative study of different machine learning methods for landslide susceptibility assessment : A case study of Uttarakhand area (India): Environmental Modelling and Software, 84, 240-250.

Pourghasemi, H.R., Rahmati, O., 2018, Prediction of the landslide susceptibility: Which algorithm, which precision ?: Catena, 162, 177-192.

Raileanu, L.E., Stoffel, K., 2004, Theoretical comparison between the Gini Index and Information Gain criterio: Annals of Mathematics and Artificial Intelligence, 2100, 77-93.

Ramos, A.M., Prada, L.F., Trujillo, M.G., Macías, J.P., Santos, A.C., 2016, Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia: Landslides, 13(4), 671-681.

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., y Guzzetti, F., 2018, A review of statistically-based landslide susceptibility models: Earth-Science Reviews, 180, 60-91.

Matich, D.J., 2001, Redes Neuronales: Conceptos Básicos y Aplicaciones: Rosario, Argentina, Universidad Tecnológica Nacional-Facultad Regional Rosario, Departamento de Ingeniería Química, reporte técnico, 55 pp.

Sabatakakis, N., Koukis, G., Vassiliades, E., Lainas, S., 2013, Landslide susceptibility zonation in Greece: Natural Hazards, 65(1), 523-543.

Sepúlveda, S.A., Petley, D.N., 2015, Regional trends and controlling factors of fatal landslides in Latin America and the Caribbean: Natural Hazards and Earth System Sciences, 15, 1821-1833.

Smith, L.I., 2002, A tutorial on Principal Components Analysis: Dunedin, University of Otago, Department of Computer Science, reporte técnico, 26 pp.

Sorbino, G., Sica, C., Cascini, L, 2010, Susceptibility analysis of shallow landslides source areas using physically based models: Natural hazards, 53(2), 313-332.

Srinivasan, K., Fisher, D., 1995, Estimating Software Development Effort: IEEE Transactions on Software Engineering, 21(2), 126-137.

van Westen, C.J., Rengers, N., Soeters, R., 2003, Use of geomorphological information in indirect landslide susceptibility assessment: Natural hazards, 30(3), 399-419.

van Westen, C.J, van Asch, T.W., Soeters, R., 2006, Landslide hazard and risk zonation — why is it still so difficult ?: Bulletin of Engineering Geology and the Environment, 65(2), 167-184.

van Westen, Cees, J., Castellanos, E., Kuriakose, S.L., 2008, Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview: Engineering Geology, 102(3-4), 112-131.

Villardón, J.L., 2002, Análisis de componentes principales: Barcelona, España, Universidad Abierta de Catalunya, Departamento de Estadística, reporte técnico, 32 pp.

Zieher, T., Schneider-Muntau, B., & Mergili, M, 2017, Are real-world shallow landslides reproducible by physically-based models? Four test cases in the Laternser valley, Vorarlberg (Austria): Landslides, 14(6), 2009-2023.

Manuscrito recibido: noviembre 18, 2020
Manuscrito corregido recibido: enero 12, 2021
Manuscrito aceptado: enero 14, 2021