Área Ambiental

Área: ambiental

Fecha de recibido: 12-04-2023
Fecha de aceptado: 22-06-2023
DOI: 10.22490/21456453.6755

PREDICCIÓN DE LA EROSIÓN DEL SUELO MEDIANTE RANDOM FOREST: CASO DE ESTUDIO CUENCA RÍO GRANDE, ANTIOQUIA

PREDICTION OF SOIL EROSION BY RANDOM OREST: CASE STUDY OF THE RIO GRANDE BASIN, ANTIOQUIA

Laura Isabel Arango-Carvajal
Ingeniera Ambiental, Magister en Bosques y Conservación Ambiental, Universidad Nacional de Colombia, Sede Medellín. Colombia. larangoc@unal.edu.co

Citación: Arango-Carvajal, L. (2024). Predicción de la erosión del suelo mediante Random Forest: caso de estudio cuenca Río Grande, Antioquia. *Revista de Investigación Agraria y Ambiental* 15(1), 317-339. https://doi.org./10.22490/21456453.6755

RESUMEN

Contextualización: actualmente, el conocimiento de fenómenos naturales asociados a la preservación de los sistemas es de interés tanto para investigadores de las ciencias naturales, como para las autoridades ambientales encargadas de la toma de decisiones sobre el manejo de los recursos. En ese sentido, se ha venido trabajando en la interpretación y predicción de diferentes fenómenos físicos como la erosión, a fin de crear escenarios que permitan fortalecer los criterios de respuesta frente a la conservación del capital natural del suelo.

Vacío de conocimiento: la capacidad de predecir el fenómeno de la erosión es limitada en muchas ocasiones, debido a la cantidad y variabilidad de los parámetros y variables que son relacionados a la erosión. Además, en muchos casos se requiere de un alto procesamiento computacional para lograr que se asocien entre sí.

Propósito: se busca implementar un modelo de machine learning como herramienta alternativa para la modelación y predicción de la erosión.

Metodología: en este estudio, se desarrolla una modelación a partir del entrenamiento del método no paramétrico Random Forest, mediante el aprendizaje supervisado, para predecir la ocurrencia de la erosión en la cuenca de Río Grande, considerando las variables que previamente han sido empleadas en otros métodos para modelar este fenómeno.

Resultados y conclusiones: los resultados mostraron una capacidad para predecir la erosión en la cuenca con una precisión aproximada del 77%, por lo que este método puede ser aplicado para obtener predicciones rápidas y confiables. Además, se encontró que las variables empleadas en el modelo RUSLE explican mayoritariamente la ocurrencia, o no, de la erosión. Finalmente, se resalta la importancia de la variable "temperatura" introducida en el modelo.

Palabras clave: machine learning, modelación ambiental, modelos no paramétricos de clasificación, aprendizaje supervisado

ABSTRACT

Contextualization: Currently, the knowledge of natural phenomena associated with the preservation of the systems is of interest both for researchers in the natural sciences, and for the environmental authorities in charge of decision-making on resource management. In this sense, work has been carried out on the interpretation and prediction of different physical phenomena such as erosion, to create scenarios that allow strengthening the response criteria for the conservation of the natural capital of the soil.

Knowledge gap: The ability to predict the phenomenon of erosion is limited on many occasions due to the quantity and variability of the parameters and variables that are related to erosion; besides that, in many cases, a high computational processing is required to achieve that they are associated with each other.

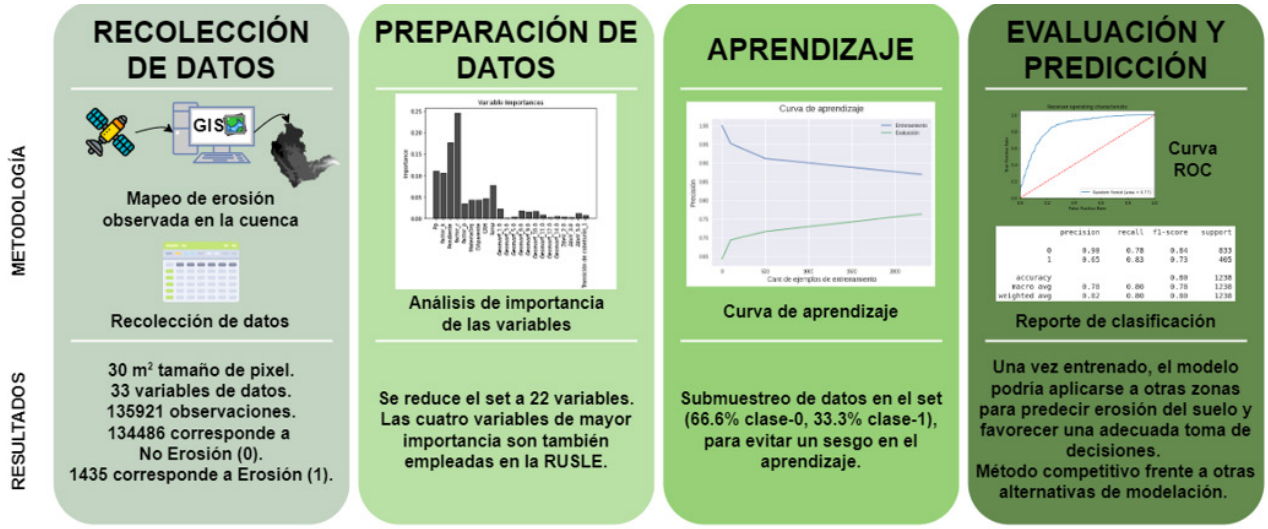
Purpose: The aim is to implement a machine learning model as an alternative tool for complex modeling and erosion prediction.

Methodology: In this study, a model is developed from the training of the non-parametric Random Forest method through supervised learning, to predict erosion occurrences in the Rio Grande basin, considering the variables that have previously been used in other methods to model this phenomenon.

Results and conclusions: The results showed a capacity to predict erosion in the basin with an approximate precision of 77%, so this method can be applied to obtain fast and reliable predictions. In addition, it was found that the variables used in the RUSLE model mainly explain the occurrence or not of erosion. The great importance of the temperature variable introduced in the model is also surprising.

Keywords: Erosion, Random Forest, non-parametric classification models, supervised learning.

RESUMEN GRÁFICO



Resumen gráfico: fases metodológicas y los resultados obtenidos en cada una de ellas
Fuente: autora.

1. INTRODUCCIÓN

La preservación de las áreas naturales es una de las estrategias más importantes para asegurar el suministro y disfrute de un flujo variado y diverso de los servicios de los ecosistemas (Martín-López *et al*., 2012). Es ahí, donde radica la importancia de planificar adecuadamente los territorios, lo que a su vez requiere de conocer el comportamiento de los sistemas, sus estados y dinámica en general (Le Clec'h *et al*., 2016). Para acercarse a una descripción más detallada, se hace necesario el establecimiento de indicadores y la recolección de datos para su cuantificación a través de monitoreos, estadísticas o modelos; de manera que, al vincular esta información a los sistemas de información geográfica, se pueda evaluar la oferta y la demanda de los recursos, y transferir los resultados a diferentes escalas espaciales y temporales (Burkhard *et al*., 2014).

Actualmente, en las ciencias naturales y en otras ciencias, es común emplear modelos simples y complejos para acercarse a la predicción de fenómenos, como por ejemplo la erosión, a partir de observaciones históricas de su comportamiento, y de su relación con determinadas variables que en conjunto se manejan como un grupo de predictores observados. Adicionalmente, se ha recurrido al estudio profundo de los fenómenos que dan criterio

a las formas de modelar, analizar y comprender cada fenómeno (Louppe, 2014).

Dentro de los métodos más conocidos para la modelación y predicción en el comportamiento de la erosión, se encuentran los mapas temáticos, puesto que tienen un alto potencial para la visualización de fenómenos complejos (Burkhard *et al*., 2014). Existen tres enfoques principales sobre los cuales se viene desarrollando este mapeo: la valoración monetaria en correspondencia a la cobertura del suelo basado en estudios previos, los métodos de valor comunitario del lugar, y las evaluaciones socioecológicas con la modelación (Martínez-Harms y Balvanera, 2012). Sin embargo, en muchas ocasiones las deducciones de mapeo sobre un fenómeno físico como la erosión y para una misma región, pueden derivar en resultados divergentes con mínimas variaciones en los métodos e incluso alejarse de la realidad.

Adicionalmente, debido a los avances tecnológicos, la recopilación de los datos correspondientes tanto a las variables predictoras como a la respuesta, se ha facilitado, de manera que la cantidad de información captada ha incrementado considerablemente, y su manipulación puede llegar a requerir un alto costo computacional (Criminisi y Shotton, 2013). En ese sentido, se han venido evaluando alternativas como la inteligencia artificial apoyada en la estadística mediante el aprendizaje supervisado y no supervisado de diversos fenómenos naturales de estudio común (Martínez-Harms y Balvanera, 2012). El desarrollo de métodos asociados a machine learning ofrecen una alternativa que, además de proveer información útil en la modelación y predicción de fenómenos naturales, ofrece ventajas sobre la capacidad de emplear un conjunto de datos de grandes dimensiones.

Random Forest es una combinación de árboles predictivos (clasificadores débiles); es decir, una modificación del Bagging (Louppe, 2014) el cual trabaja con una colección de árboles de decisiones intercorrelacionadas, y los promedia. En ese sentido, es considerado uno de los mejores algoritmos de clasificación, capaz de clasificar grandes cantidades de datos con precisión, considerando múltiples variables, ya que selecciona submuestras para elaborar cada árbol (Medina-Merino y Ñique-Chacón, 2017). En el presente estudio se analiza la capacidad de predicción de la erosión en la cuenca hidrográfica de estudio Río Grande, en el departamento de Antioquia (Andes colombianos), a partir de la aplicación del modelo Random Forest. Para ello, se han contemplado variables climatológicas, edáficas y de cobertura vegetal, que comúnmente son empleadas para predecir este fenómeno; además de un mapa de erosión construido a partir de imágenes satelitales de la zona de estudio como variable de respuesta.

La metodología planteada se encuentra encaminada a enriquecer las posibilidades de modelación de la comunidad científica en alianza con las entidades públicas encargadas del ordenamiento y

planificación territorial para la toma de decisiones sobre los ecosistemas, en la medida en que puede madurarse para llegar a ser una herramienta ágil con buenos resultados.

2. MÉTODOS

2.1. Zona de estudio

La cuenca de Río Grande se encuentra ubicada en el altiplano norte del departamento de Antioquia, en los Andes colombianos (Figura 1). Su nacimiento tiene lugar en el municipio de Santa Rosa de Osos a los 3364 m s. n. m. aproximadamente, y comprende los municipios de Entrerríos, Belmira, San Pedro de los Milagros y Don Matías, en donde desemboca al río Medellín a una altitud de 1068 m s. n. m. para dar lugar a la formación del río Porce (Corantioquia y UNAL, 2015).

La cuenca tiene una extensión aproximada de 1280 km^2. La precipitación media anual en la región es de 2238 mm, distribuida en un régimen bimodal con dos temporadas húmedas (abril-mayo y octubre-noviembre) y dos temporadas secas, (diciembre-febrero y junio-agosto) junto con una humedad relativa del 79% (Suescún *et al*., 2017). En la zona de estudio, predomina el clima frío húmedo y frío muy húmedo, presentándose una temperatura promedio de 14°C (Machado *et al*., 2019)the literature on ecosystem services has pointed to the need to quantify and characterise Soil Natural Capital (SNC. En esta región, la topografía es ondulada de colinas suaves, los suelos son en su mayoría derivados de ceniza volcánica, de baja fertilidad, muy ácidos, con deficiencia en fósforo y calcio, y con un contenido de materia orgánica media a alta (Suescún *et al*., 2017). La zona de vida en la región corresponde al bosque muy húmedo montano bajo (bmh-MB) (Ramírez *et al*., 2018).

En términos ecológicos es una región de interés, ya que el uso del suelo se ve determinado por el desarrollo de actividades económicas como la ganadería de leche, que ha desencadenado la siembra de pasto sobre una extensión significativa de su territorio, sumado a los cultivos de papa, tomate de árbol y algunos productos de pancoger (García-Leoz *et al*., 2018); por otro lado, la intensificación del recurso hídrico para el aprovechamiento de agua en la generación de energía eléctrica, ya que allí se encuentran los embalses de Quebradona, y Río Grande, los cuales proveen agua potable a cerca de dos millones de habitantes en el Valle de Aburrá (Suescún *et al*., 2018).

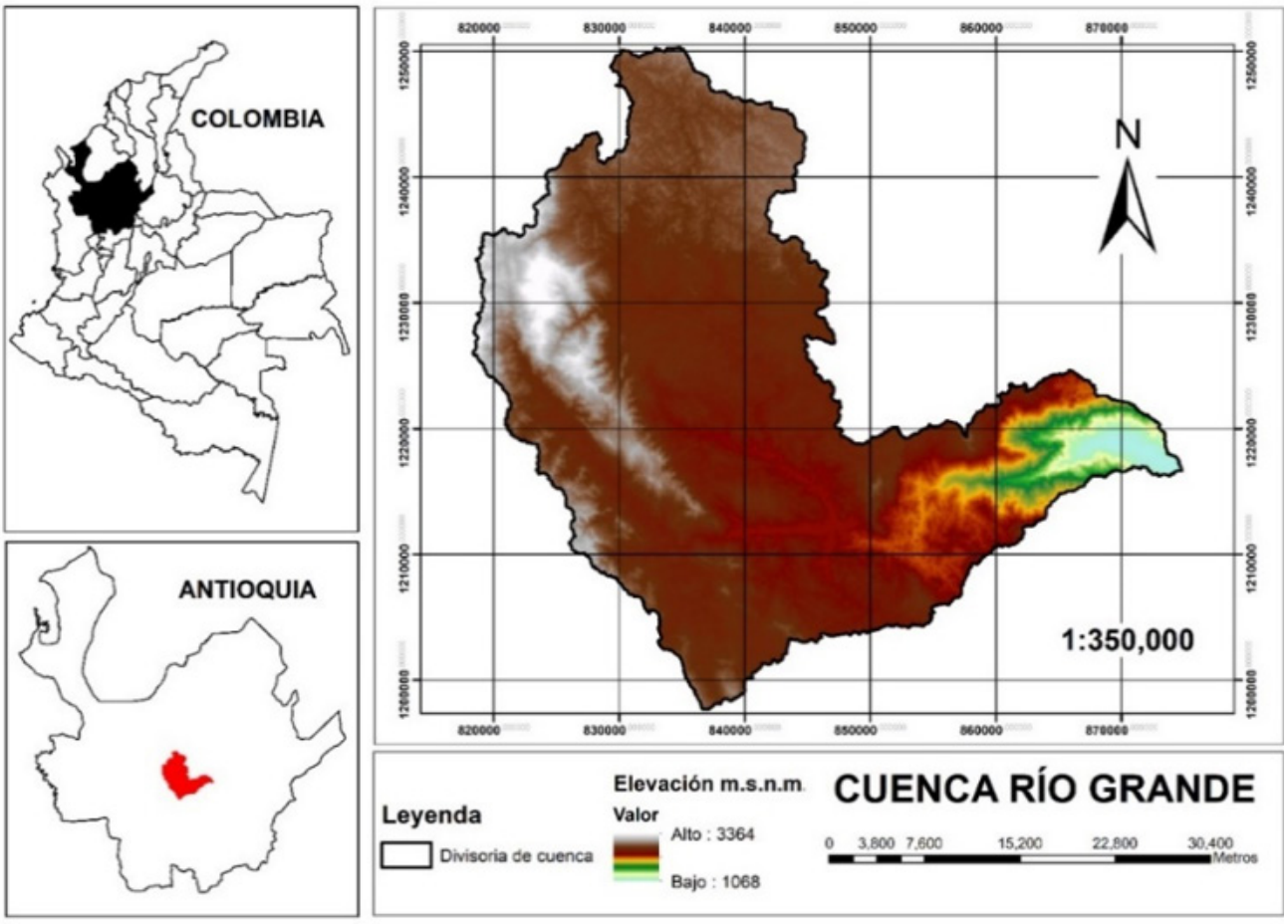


Figura 1. Cuenca de Río Grande.
Fuente: autora.

2.2. Insumos

La erosión es un fenómeno físico ampliamente estudiado, principalmente en las regiones donde su ocurrencia puede estar asociada a la pérdida de servicios ambientales (Sepúlveda, 2013). En la cuenca de Río Grande en Antioquia, se conoce el desarrollo de dos trabajos de modelación que han procurado predecir la erosión bajo diferentes escenarios. De estos ejercicios, se tomaron las variables empleadas para cada caso a partir de mapas espacializados para toda la cuenca en formato ráster (Anexo 1), que a continuación se han convertido en archivos tipo vector para formar una matriz o set de datos con la información inicial. Adicionalmente, para complementar el ejercicio, se introdujeron las variables de temperatura, transición de coberturas, geomorfología y zonas de vida del área de estudio en el set de datos inicial (Anexo 2).

El primer ejercicio de modelación, considera la ecuación empírica para la pérdi-

da de suelos propuesta por Wischmeier y Smith (1978), que define una serie de variables con base en la precipitación del área de estudio, las propiedades del suelo y su uso, su topografía y las coberturas vegetales con diferentes manejos (Ecuación 1) (Mohammed *et al*., 2020). Este ejercicio corresponde a la tesis de maestría *Modelo para la definición de áreas estratégicas para la conservación de suelos a partir de la susceptibilidad a la erosión hídrica* (Sepúlveda, 2013).

$$E=R*K*LS*C*P \quad \text{(Ecuación 1)}$$

Donde, E es la Erosión hídrica, R es Factor de erosividad de la lluvia, K es el Factor de erodabilidad del suelo, LS es el Factor de longitud de la pendiente del terreno, C es el Factor de cobertura y manejo, y P es el Factor de las prácticas de soporte. En este caso las variables consideradas en el set de datos corresponden a la precipitación (Pp), Factor K, pendiente, Factor C y Factor P (Anexo 2).

El segundo ejercicio considera que la vulnerabilidad a la erosión depende únicamente de las variables (propiedades) del suelo tales como materia orgánica, densidad aparente, capacidad de retención de la humedad y pendiente; y que en su conjunto como capital natural provee servicios potencialmente aprovechables por la comunidad (Machado *et al*., 2019) the literature on ecosystem services has pointed to the need to quantify and characterise Soil Natural Capital (SNC. La información correspondiente a este estudio pertenece a la tesis de maestría *Impacto potencial de pérdida del servicio ecosistémico intermedio de control de erosión por cambios en el capital natural del suelo. Caso de estudio: Cuenca de Río Grande, Departamento de Antioquia* (Machado, 2018).

$$CVI=0.3(SOM) + 0.1(BD) + 0.2(AWHC) + 0.4(S) \quad (\textit{Ecuación 2})$$

Donde, CVI es el Índice de vulnerabilidad, SOM es la Materia orgánica del suelo, BD es la Densidad aparente, AWHC es la Capacidad de almacenamiento de agua y S es la Pendiente del terreno. Estas cuatro variables (propiedades del suelo) fueron también consideradas en el set de datos (Anexo 2).

Adicionalmente, a través de la herramienta *Google Earth Pro* y la combinación de imágenes de alta resolución como referencia, se generaron polígonos contemplando todas las zonas en donde se pudieran visualizar procesos de erosión (UPRA, 2021). Una vez obtenidos todos los polígonos se unificaron en una categoría y se llevaron a formato ráster. De esta manera, se obtuvo un mapa denominado "Erosión observada" que también se convirtió en archivo tipo vector para completar el set de datos (Anexo 3). Esta información corresponde a la variable respuesta para el entrenamiento del modelo.

2.3. Implementación del algoritmo Random Forest

2.3.1 Preparación del set de datos

Este es un ejercicio de clasificación binaria que se desarrolló con el objetivo de predecir si en un territorio se presenta erosión,

o no (1 y 0 respectivamente – "erosión observada"). Para ello, se parte del set de datos mencionado en el apartado anterior que comprende las variables típicamente usadas en el modelamiento del fenómeno de la erosión. El desarrollo del análisis de los datos se realizó usando el lenguaje de programación *Python* en la versión 3.9. El primer tratamiento que se le realizó a los datos consistió en eliminar todas las observaciones que tuvieran datos faltantes de manera que se pudiera asegurar que todos los datos contaban con la información de cada variable considerada. Se identificó que ninguna de las variables superaba el 10% de datos faltantes.

Posteriormente, se analizó la representación numérica de cada una de las variables categóricas, encontrando que únicamente las variables de transición de coberturas, geomorfología y zonas de vida estaban sobredimensionadas, lo que podría ocasionar problemas debido a la jerarquía o peso asignado. Por esta razón, se crearon variables binarias para cada uno de los factores "Transición de coberturas", "Geomorf" y "ZdeV". De esta manera, el set de datos pasó de tener 12 variables a 33. A continuación, para facilitar el desempeño de los modelos, se procedió a escalar cada una de las variables con la función de preprocesamiento de datos *MinMaxScaler* de la librería *sklearn* (Hao y Ho, 2019). Finalmente, se hizo una selección de variables considerando la correlación entre ellas y su importancia en la explicación del problema (Anexo 4).

2.3.2. Diseño del modelo Random Forest

Para un primer ejercicio de estimación del desempeño del modelo, se empleó el set de datos mencionado anteriormente. El modelo implementado se basó en el algoritmo de ensamble Random Forest de clasificación de la librería *sklearn*, el cual crea un conjunto de árboles de decisión a partir de un subconjunto de datos de entrenamiento seleccionados al azar. Cada árbol de decisión en Random Forest realiza una predicción de una clase, de manera que la clase con más votos se convierte en la predicción de nuestro modelo (Genuer y Poggi, 2020).

Inicialmente, el algoritmo se desarrolló con un número de árboles 'n_estimatiors' preestablecido en 100, entendiendo que mientras más grande sea esta variable se obtienen mejores resultados; sin embargo, se aumenta la carga computacional y dejará de presentar una mejora significativa cuando alcance un número crítico de árboles. El tamaño de los subconjuntos aleatorios de las características '*max_features*' fue inicializado en *"sqrt(n_features)"* donde *n_features* es el número de características, considerando que mientras menor sea esta variable, mayor será la reducción de la varianza, pero también aumentará el sesgo. La profundidad máxima del árbol '*max_depth*' fue definida como *'None'*, lo cual indica que los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos de un número mínimo de muestras *('min_samples_split')* requeridas para dividir un nodo interno el cual se estableció en "2".

Se definió un criterio de índice '*gini*' como función para medir la calidad de una división (*'criterion'*), un set de validación "*hold out*" y una métrica de precisión del modelo, calculada mediante validación cruzada con cinco divisiones (Anexo 4).

2.3.3 Entrenamiento del modelo Random Forest

El entrenamiento del modelo se realizó con un 70 % de los datos obtenidos en los dos grupos 0 y 1, para un tamaño de lote de 135921 muestras. La precisión del modelo se observó mediante validación cruzada; sin embargo, la precisión definitiva de los modelos se calculó por medio de la comparación entre la predicción resultante y el 30% de los datos restantes de la variable dependiente.

2.3.4 Optimización del modelo Random Forest

- **Desbalanceo de datos**

Se identificó inicialmente un problema de desbalanceo de datos, debido a que la mayoría de los datos pertenecen a la clase 0 —no erosión (99%), mientras que la clase 1— erosión corresponde al 1% de los datos. Como los conjuntos de datos no tienen valores nulos y ya están escalados, no se realizó ningún procesamiento adicional. La Figura 2 muestra el desequilibrio de datos. Para manejar este problema, se usó la técnica de asignación de pesos: el algoritmo Random Forest tiene la tendencia a estar sesgado en la clase mayoritaria, por lo tanto, se impuso una penalización a la clasificación errónea de la clase minoritaria. Los pesos de las clases se incorporaron al algoritmo y se determinaron a partir de la relación entre el número de conjunto de datos en la clase 0 y el número del conjunto de datos en la clase 1.

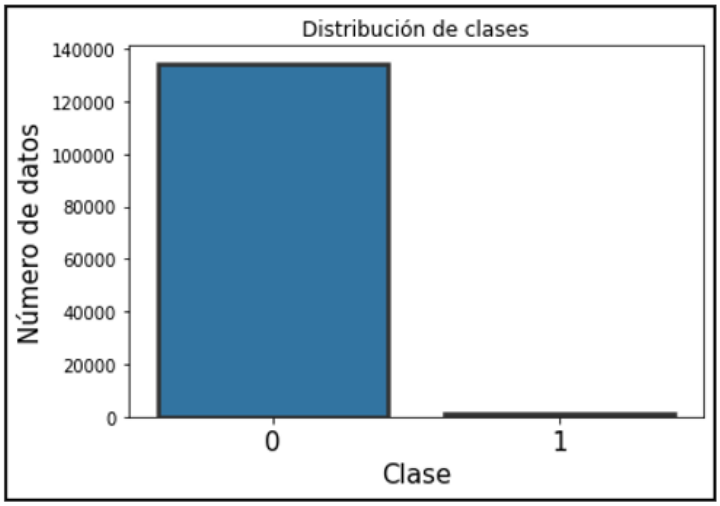


Figura 2. Reporte de clasificación entrenamiento del modelo.

Fuente: autora.

- **Muestreo para obtener datos más equilibrados**

Una técnica para balancear los datos es el remuestreo. Para el set de datos definido se realizó un muestreo descendente (submuestreo) de la clase mayoritaria, es decir, que se redujo la cantidad de muestras de la clase 0 hasta una proporción 2 a 1, lo que indica que el nuevo set de datos estará conformado por un 66.6% de muestras clase 0 y un 33.3% de muestras clase 1, con un total de 4322 muestras.

- **Afinamiento de los hiperparámetros**

En general, un hiperparámetro es un parámetro del modelo que se establece antes del inicio del proceso de aprendizaje. En el caso de Random Forest existen varios hiperparámetros diferentes que se pueden ajustar, y de diferentes formas, para elegirlos (Genuer y Poggi, 2020). En este caso, se empleó la curva de validación que es una buena forma de verificar visualmente los valores potencialmente optimizados de los hiperparámetros. Es importante tener en cuenta que, al construir las curvas de validación, los otros parámetros se mantienen en sus valores predeterminados (Anexo 5).

Otra forma de elegir qué hiperparámetros ajustar, es realizando una búsqueda aleatoria con *Random Search* o una búsqueda exhaustiva con *Grid Search*, estos métodos permiten definir rangos u opciones para realizar múltiples combinaciones de hiperparámetros, así como todas las validaciones cruzadas con el objetivo de encontrar los hiperparámetros que arrojen una mejor precisión.

Para el método Random Search (Probst *et al*., 2019) se incluyeron los siguientes hiperparámetros con sus respectivos rangos. Este método elige valores aleatorios dentro de los rangos especificados:

- N_estimators: [int(x) for x in np.linspace(start = 10, stop = 2000, num = 10)]
- N_splitmax_features = ['auto', 'sqrt']
- Max_depth: [int(x) for x in np.linspace(10, 100, num = 10)]max_depth.append(None)
- Min_samples_split: [2, 5, 3, 10]
- Min_samples_leaf: [1, 2, 3, 4]
- Bootstrap: [True, False]

Grid Search efectúa todas las combinaciones de hiperparámetros posibles con las opciones indicadas (Probst *et al*., 2019). Para este método se definieron las siguientes opciones de hiperparámetros:

- Bootstrap: [True],
- Max_depth: [5, 7, 8, 9, 10, 11, 12, 15],
- Min_samples_leaf: [3, 4, 5, 6, 7],
- Min_samples_split: [2, 3, 4, 5, 7],
- N_estimators: [1000, 1500, 1778, 1800, 2000]

- **Métricas de validación**

Como métrica de validación se empleó la curva de aprendizaje (Gómes y Carmona, 2022) que permite visualizar el efecto del número de observaciones en el desempe-

ño del modelo, graficando la precisión en función del tamaño de los datos de entrenamiento. Se utilizó con los datos de entrenamiento y con los datos de validación para determinar si el modelo se subajusta o sobreajusta a los datos. Por otro lado, también se implementó la curva de características operativas del receptor (ROC) (Gómes y Carmona, 2022), como una herramienta comúnmente utilizada con clasificadores binarios. La línea punteada representa la curva ROC de un clasificador puramente aleatorio; un buen clasificador se mantiene lo más alejado posible de esa línea (hacia la esquina superior izquierda).

3. RESULTADOS Y DISCUSIÓN

3.1. Entrenamiento y optimización del algoritmo Random Forest

De las 33 variables iniciales consideradas en el set de datos original, se encontró que 22 variables explicaban el problema en un 98%, por lo cual fueron consideradas para set inicial que corresponden a 135921 observaciones, de las cuales 134486 corresponden a No Erosión (0) y 1435 a Erosión (1).

Como resultado inicial del entrenamiento, se encuentran las métricas de clasificación consignadas en las Tablas 1 y 2.a, y una media para el score de la validación cruzada de 0.989.

Tabla 1. Reporte de clasificación entrenamiento del modelo

	precision	recall	f1-score	support
0	0.99	1.00	0.99	40330
1	0.41	0.02	0.05	447
accuracy			0.99	40777
macro avg	0.70	0.51	0.52	40777
weighted avg	0.98	0.99	0.98	40777

Fuente: autora.

Tabla 2. Matrices de confusión

a. Entrenamiento del modelo

	Clase observada negativa	Clase observada positiva
Clase observada negativa	40314	16
Clase observada positiva	436	11

b. Optimización por class_weigth

	Clase observada negativa	Clase observada positiva
Clase observada negativa	28574	11756
Clase observada positiva	59	388

c. Entrenamiento del modelo en el remuestreo de datos desequilibrados

	Clase observada negativa	Clase observada positiva
Clase observada negativa	714	119
Clase observada positiva	130	275

d. Optimización – grid search

	Clase observada negativa	Clase observada positiva
Clase observada negativa	712	121
Clase observada positiva	119	286

e. Modelo optimizado final

	Clase observada negativa	Clase observada positiva
Clase observada negativa	652	181
Clase observada positiva	70	335

Fuente: autora.

Desbalanceo de datos

Debido a la asignación de pesos realizada para compensar el desbalanceo de datos, que contaba con una proporción de aproximadamente 94 entre el número de conjuntos de datos en la clase 0 y la clase 1 (134486/1435), se desarrolló un nuevo entrenamiento donde se obtuvieron las métricas de clasificación consignadas en las Tablas 3 y 2.b, además de una media para el score de la validación cruzada de 0.719.

Tabla 3. Reporte de clasificación de la optimización por class_weigth

	precision	recall	f1-score	support
0	1.00	0.71	0.83	40330
1	0.03	0.87	0.06	447
accuracy			0.71	40777
macro avg	0.51	0.79	0.45	40777
weighted avg	0.99	0.71	0.82	40777

Fuente: autora.

Es así como se hace evidente que algunas de las herramientas comúnmente empleadas para optimizar los modelos de machine learning pueden presentar contraprestaciones. En este caso, por ejemplo, la técnica de asignación de pesos para las clases permitió obtener un aumento en la predicción de la clase 1 (erosión), sin embargo, incrementó la cantidad de falsos positivos tal como se evidenció en la matriz de confusión (Tabla 2.b).

Muestreo para obtener datos más equilibrados

Posteriormente, a partir del remuestreo del set de datos, se obtuvieron las métricas de clasificación de las Tablas 4 y 2.c, junto con una media para el score de la validación cruzada de 0.76 con el nuevo entrenamiento.

Tabla 4. Reporte de clasificación de la optimización por class_weigth

Reporte de clasificación - SD3:

	precision	recall	f1-score	support
0	0.85	0.86	0.85	833
1	0.70	0.68	0.69	405
accuracy			0.80	1238
macro avg	0.77	0.77	0.77	1238
weighted avg	0.80	0.80	0.80	1238

Fuente: autora.

Afinamiento de los hiperparámetros

En la Tabla 5, se presentan los reportes para la combinación de hiperparámetro. Se encontró que el mayor accuracy está dado por el método grid search, y se obtienen las métricas de clasificación de las Tabla 6 y 2.d, junto con una media para el score de la validación cruzada de 0.766.

Tabla 5. Hiperparámetros

Hiperparámetro	Valor por defecto	Grid search	Randomiz Search
Bootstrap	True	True	True
Max_depth	None	8	10
Min_samples_ leaf	1	5	4
Min_samples_ split	2	2	3
N_estimators	100	1778	1778
Accuracy	0.7668	0.7730	0.7648

Fuente: autora.

Tabla 6. Reporte de clasificación de la optimización-grid search

	precision	recall	f1-score	support
0	0.86	0.85	0.86	833
1	0.70	0.71	0.70	405
accuracy			0.81	1238
macro avg	0.78	0.78	0.78	1238
weighted avg	0.81	0.81	0.81	1238

Fuente: autora.

3.2. Análisis de características con Random Forest

Como proceso en paralelo, se realizó un análisis de características importantes entre las 22 variables mediante el algoritmo de Random Forest; posterior al entrenamiento del modelo, se usaron los porcentajes relativos devueltos por el modelo para clasificar las características según su importancia. En ese sentido se puede ver que las variables más importantes corresponden al factor C de la cobertura vegetal, seguida de la pendiente del terreno, la precipitación y el factor K de la erodabilidad; dentro de las variables introducidas, se observa que la temperatura presenta la mayor importancia y que las demás variables poco explican el modelo.

En suma, cabe anotar que las variables empleadas en el cálculo del IVR (materia orgánica, densidad aparente y capacidad de retención de humedad) tiene una importancia similar a la hora de explicar el modelo, por debajo de las variables asociadas al cálculo de la RUSLE en la zona de estudio.

Es necesario evaluar el comportamiento de la importancia de las variables para explicar el problema de la predicción de la erosión en la cuenca, ya que permite entender cuál de las aproximaciones en la modelación puede detectar mejor este fenómeno. En este caso se encontró que de las cinco variables empleadas en el método de la ecuación universal de la pérdida de suelo ajustada RUSLE, cuatro presentan la mayor importancia para explicar el fenómeno de la erosión, aun cuando este método es ampliamente criticado por las generalidades que acarrea su aplicación en los paisajes tropicales (Barral, 2016).

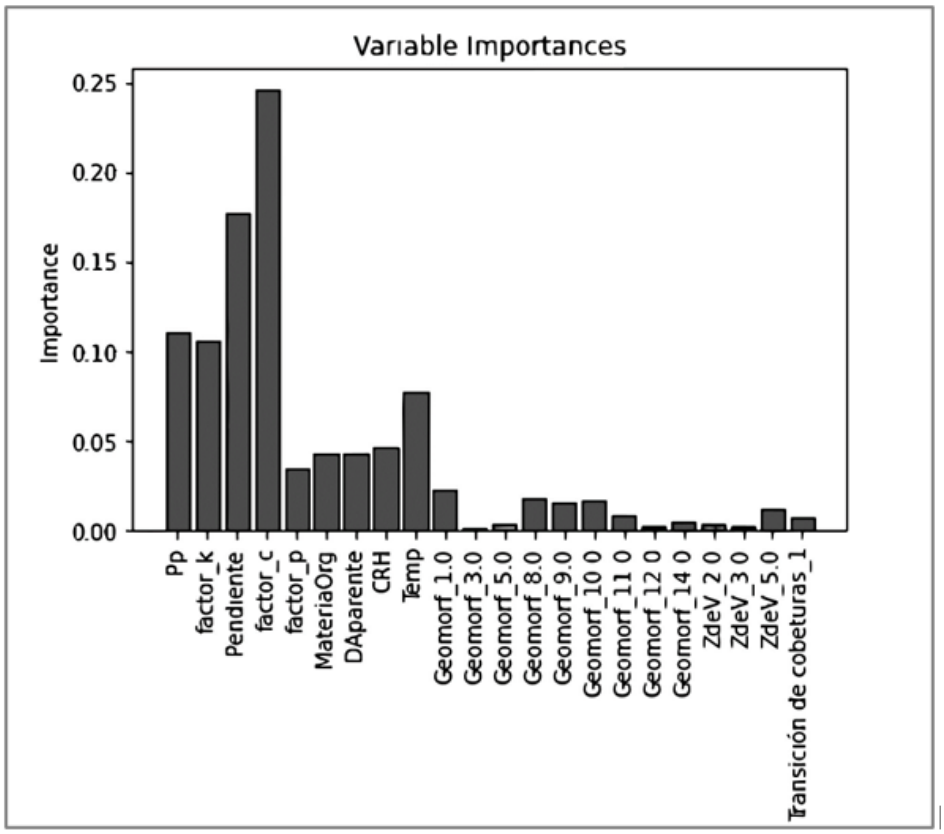


Figura 3. Importancia de las variables.
Fuente: autora.

Por otro lado, sorprende la baja importancia asignada a la variable introducida de la transición de las coberturas (Figura 3), ya que teóricamente esta es una de las variables que más incidencia tienen en el fenómeno de la erosión del suelo (Duan *et al*., 2023; Wang *et al*., 2022). Asimismo, la temperatura que se esperaba tuviese una menor relación con la explicación del fenómeno, presentó una importancia incluso mayor que las variables empleadas en el método del IVR.

3.3 Modelo optimizado

Mediante los procesos de optimización del modelo Random Forest para la predicción de la susceptibilidad a la erosión, a través de grid_search se obtuvieron hiperparámetros que arrojaron una mejor precisión en la métrica de validación cruzada; adicionalmente, se aplicó class_weigth para obtener el modelo más optimizado. Lo anterior, se presenta en el reporte de clasificación y la matriz de confusión (Tabla 7). La media del score de cada una de las divisiones de la validación cruzada corresponde a 0.769 (Tabla 7).

Tabla 7. Reporte de clasificación para el modelo optimizado

	precision	recall	f1-score	support
0	0.90	0.78	0.84	833
1	0.65	0.83	0.73	405
accuracy			0.80	1238
macro avg	0.78	0.80	0.78	1238
weighted avg	0.82	0.80	0.80	1238

Fuente: autora.

De estos resultados se observa que el modelo aprendió de manera eficiente a identificar el fenómeno de erosión en la cuenca, sin embargo, respecto a los entrenamientos anteriores, disminuyó considerablemente la capacidad de acertar en la predicción de los valores donde no se presenta erosión. Aun así, presenta la mejor matriz de confusión al igual que la mejor media del score entre cada uno de los subconjuntos de la validación cruzada.

En general, se evidencia que el ejercicio de entrenamiento de modelos con datos desbalanceados, ofrece un reto a superar debido a que generalmente se presentan problemas para detectar las respuestas de la clase que presenta menor proporción (Cárdenas, 2019). A pesar de arrojar una precisión muy elevada, el modelo presenta un sobreentrenamiento de datos de la clase mayoritaria que lo entorpece para la predicción de la clase minoritaria. En nuestro caso de estudio, el modelo entrenado con datos desequilibrados pierde valor debido a que se desea predecir la erosión en la cuenca y, con dichas condiciones en el set de datos, se entorpece la predicción acertada de estos datos de interés.

3.4. Curva de aprendizaje y curva ROC

Este resultado se complementa con el resultado de la curva de validación (Anexo 5), que muestra una tendencia, del modelo, a seguir aprendiendo. En el caso de la curva ROC, es evidente que, si bien la línea azul se mantiene alejada de línea roja, no tiene la pendiente deseada o no se acerca totalmente a la esquina superior izquierda lo que permite inferir que si bien este modelo es una herramienta buena de clasificación, todavía presenta oportunidades de mejora.

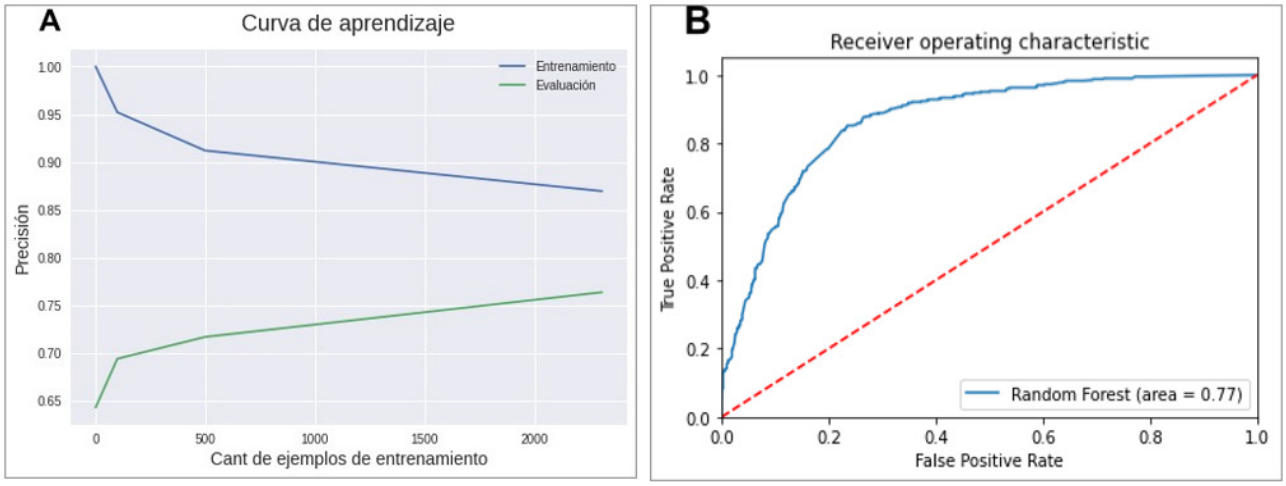


Figura 4. A. Curva de aprendizaje y B. Curva ROC.
Fuente: autora

4. CONCLUSIONES

Este método puede ser aplicado para obtener predicciones rápidas y confiables de la ocurrencia del fenómeno de la erosión, lo que posibilita una eficaz tom de decisiones sobre el manejo de los suelos, inclusive, a partir del análisis de las variables evaluadas (Barrera *et al*., 2013; Suescún *et al*., 2017). Si bien se esperaban mejores resultados en el desarrollo del ejercicio planteado, aplicar el modelo Random Forest para predecir la erosión en la cuenca de Río Grande es competitivo frente a otras alternativas de modelación y mapeo de fenómenos naturales que requieren mayor costo computacional, además de limitaciones en el uso de datos y de variables.

Estudios futuros basados en la metodología acá presentada podrían considerar la inclusión de otras variables asociadas al suelo que puedan explicar con mayor precisión el comportamiento de la erosión; así como la metodología para el mapeo de los procesos erosivos reales, podría darse mediante metodologías como el cálculo del NDVI, de manera que se amplíe el espectro de los procesos erosivos de menor escala, y a su vez, se logren datos más equilibrados en las clases de interés de erosión y no erosión. Lo anterior considerando que mediante el método aquí presentado se obtuvieron áreas significativamente pequeñas con presencia de erosión, lo cual dificulta el aprendizaje del modelo.

Finalmente, el análisis de importancia de variables se convierte en una herramienta fundamental, sobre todo para definir las variables con más importancia a la hora de explicar el fenómeno de la erosión,

pero también porque permite considerar variables que poco se asocian al fenómeno. Tal es el caso de la temperatura, que se identificó como una característica relevante, lo cual sugiere que esta variable debería ser tenida en cuenta para mejorar la precisión de los modelos, y las ecuaciones de susceptibilidad y vulnerabilidad a la erosión existentes en la actualidad.

CONTRIBUCIÓN DE LA AUTORÍA

Laura Isabel Arango-Carvajal: metodología, conceptualización, investigación, análisis de datos, escritura y edición.

AGRADECIMIENTOS

Agradecimientos al profesor Edier Aristizábal, de la Facultad de Minas de las Universidad Nacional de Colombia, sede Medellín, por el conocimiento brindado que hizo posible el desarrollo del presente proyecto. Igualmente, al ingeniero de telecomunicaciones Juan David Chimá por sus aportes a la programación del modelo y análisis de datos. Finalmente, a Colciencias, por proveer los recursos económicos para esta investigación a través del Proyecto "Trayectorias de sistemas socio-ecológicos y sus determinantes en cuencas estratégicas en un contexto de cambio ambiental. Código 110180863961" Convocatoria 808-2018 Proyectos de ciencia, tecnología e innovación y su contribución a los retos de país.

LITERATURA CITADA

Barral, M. P. (2016). *Tutorial para el mapeo de funciones ecosistémicas y servicios ecosistémicos con protocolo ECOSER* (Vol. 1). Unidad Integrada Balcarce (EEA INTA Balcarce – Facultad de Ciencias Agrarias, UniversidadNacional de Mar del Plata

Barrera, J. E., Rivera, J. H., y Cadena, M. E. (2013). Influencia del sistema radical de cuatro especies vegetales en la estabilidad de laderas a movimientos masales. *Cenicafé*, *64*(2), 59-77. http://biblioteca.cenicafe.org/bitstream/10778/531/1/arc064%2802%2959-77.pdf

Burkhard, B., Kandziora, M., Hou, Y., & Müller, F. (2014). Ecosystem service potentials, flows and demands-concepts for spatial localisation, indication

and quantification. *Landscape Online*, *34*(1), 1-32. https://doi.org/10.3097/LO.201434

Cárdenas, J. A. (2019). *Clasificación de aceptación de campañas para una entidad financiera, usando random forest con datos balanceados y datos no balanceados*. [Tesis de Maestría]. Universidad Ricardo Palma Escuela de Posgrado.

Corantioquia y UNAL. (2015). *Actualización y ajuste Plan de Ordenación y Manejo de la Cuenca de los ríos Grande y Chico.* Contrato 967 de 2013.

Criminisi, A., & Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4471-4929-3

Duan, X., Chen, Y., Wang, L., Zheng, G., & Liang, T. (2023). The impact of land use and land cover changes on the landscape pattern and ecosystem service value in Sanjiangyuan region of the Qinghai-Tibet Plateau. *Journal of Environmental Management*, *325*(PB), 116539. https://doi.org/10.1016/j.jenvman.2022.116539

García-Leoz, V., Villegas, J. C., Suescún, D., Flórez, C. P., Merino-Martín, L., Betancur, T., & León, J. D. (2018). Land cover effects on water balance partitioning in the Colombian Andes: improved water availability in early stages of natural vegetation recovery. *Regional Environmental Change*, *18*(4), 1117-1129. https://doi.org/10.1007/s10113-017-1249-7

Genuer, & Poggi, J. (2020). Random Forests with R. In *Use R*. https://doi.org/10.1007/978-3-030-56485-8

Gómes, N., & Carmona, M. (2022). An application of Machine learning Techniques to the Prediction of Purchase in the Tourism Sector. *EasyChair*.

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348-361. https://doi.org/10.3102/1076998619832248

Le Clec'h, S., Oszwald, J., Decaens, T., Desjardins, T., Dufour, S., Grimaldi, M., Jegou, N., & Lavelle, P. (2016). Mapping multiple ecosystem services indicators : Toward an objective-oriented approach. *Ecological Indicators*, *69*, 508-521. https://doi.org/10.1016/j.ecolind.2016.05.021

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. *July*. http://arxiv.org/abs/1407.7502

Machado, J. (2018). *Impacto Potencial de pérdida del servicio ecosistémico intermedio de control de erosión por cambios en el capital natural del suelo. Caso de estudio: Cuenca de Riogrande, Departamento de Antioquia* [Universidad Nacional de Colombia]. https://doi.org/10.1016/j.ecolind.2017.07.051

Machado, J., Villegas-Palacio, C., Loaiza, J. C., & Castañeda, D. A. (2019). Soil natural capital vulnerability to environmental change. A regional scale

approach for tropical soils in the Colombian Andes. *Ecological Indicators*, *96*(May 2018), 116-126. https://doi.org/10.1016/j.ecolind.2018.08.060

Martín-López, B., González, J., & Vilardy, S. (2012). *Guía Docente Ciencias de la Sostenibilidad*. CO-BAC. https://doi.org/10.1016/j.ajhg.2011.11.018

Martínez-Harms, M. J., & Balvanera, P. (2012). Methods for mapping ecosystem service supply : a review. *International Journal of Biodiversity Science, Ecosystem Services & Management*, *8*(1-2), 17-25. https://doi.org/10.1080/21513732.2012.663792

Medina-Merino, R. F., y Ñique-Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*, *0*(010), 165. https://doi.org/10.26439/interfases2017.n10.1775

Mohammed, S., Alsafadi, K., Talukdar, S., Kiwan, S., Hennawi, S., Alshihabi, O., Sharaf, M., & Harsanyie, E. (2020). Estimation of soil erosion risk in southern part of Syria by using RUSLE integrating geo informatics approach. *Remote Sensing Applications: Society and Environment*, *20*(2019), 100375. https://doi.org/10.1016/j.rsase.2020.100375

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Ramírez, C. D., Orrego, S. A., & Schneider, L. C. (2018). Identifying Drivers and Spatial Patterns of Deforestation in the Río Grande Basin, Colombia. *Journal of Latin American Geography*, *17*(1), 108-138. https://doi.org/10.1353/lag.2018.0005

Sepúlveda, L. (2013). *Modelo para la definición de áreas estratégicas para la conservación de suelos a partir de la determinación de la susceptibilidad a la erosión hídrica*. Universidad de Antioquia.

Suescún, D., León, J. D., Villegas, J. C., García-Leoz, V., Correa-Londoño, G. A., & Flórez, C. P. (2018). ENSO and rainfall seasonality affect nutrient exchange in tropical mountain forests. *Ecohydrology*, *e2056*, 1-10. https://doi.org/10.1002/eco.2056

Suescún, D., Villegas, J. C., León, J. D., Flórez, C. P., García-Leoz, V., & Correa-Londoño, G. A. (2017). Vegetation cover and rainfall seasonality impact nutrient loss via runoff and erosion in the Colombian Andes. *Regional Environmental Change*, *17*(3), 827-839. https://doi.org/10.1007/s10113-016-1071-7

UPRA, U. de P. R. A.-. (2021). Evaluaciones Agropecuarias Municipales – EVA. Módulo de Consulta de Información. https://www.upra.gov.co/web/guest/consulta-de-informacion

Wang, P., Li, R., Liu, D., & Wu, Y. (2022). Dynamic characteristics and responses of ecosystem services under land use/land cover change scenarios in the Huangshui River Basin, China. *Ecologi-*

cal Indicators, *144*(May), 109539. https://doi.org/10.1016/j.ecolind.2022.109539

Wischmeier, W. H., & Smith, D. D. (1978). Predicting rainfall erosion losses: a guide to conservation planning. Department of Agriculture, Science and Education Administration.

ANEXOS

Anexo 1. Mapas ráster de las variables involucradas en la modelación

Anexo 2. Set de datos inicial

Anexo 3. Mapa de erosión observada

Anexo 4. Scripts empleados

Anexo 5. Curva de validación para class_weight