



EVALUACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE BORO EN SUELOS ANDISOLES DE NARIÑO-COLOMBIA †

[EVALUATION OF MACHINE LEARNING MODELS FOR BORON PREDICTION IN ANDISOL SOILS OF NARIÑO-COLOMBIA]

D. E. Álvarez-Sánchez¹, X. López-Estrella², E. Manso-Ordoñez²,
L. López-Rivera³ and J. Rodríguez-Valenzuela¹

¹Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA, Centro de investigación Obonuco. Km 5 vía Pasto-Obonuco, Nariño, Pasto, Colombia. Email: dalvarez@agrosavia.co, jrodriguezv@agrosavia.co

²Universidad de Nariño, Facultad de Ciencias Agrícolas, Departamento de Producción y Sanidad Vegetal. Calle 18 Cr 50 Ciudadela Universitaria Torobajo. Nariño, Pasto, Colombia. Email: xilelopez@gmail.com, 94rique@gmail.com

³Universidad de Nariño, Facultad de Ciencias Exactas y Naturales, Departamento de Biología, Calle 18 Cr 50 Ciudadela Universitaria Torobajo. Nariño, Pasto, Colombia. Email: lauramlopezrivera@gmail.com

*Corresponding author

SUMMARY

Background. The southern region of Nariño in Colombia has a strong agricultural vocation. However, it has been identified that the technology associated with the recommendation of fertilization especially in microelements in the soil is limited, which generates conflicts in production. This situation demands a different and innovative approach to address these challenges. That is why this research uses the predictive capacity of the machine learning (ML) approach. **Objective.** To explore the application of ML tools for the prediction of Boron levels in Andisols soils of Nariño has been explored and identifying the most efficient algorithm. **Methodology.** A total of 1,067 soil samples collected in various fields of five municipalities in the southern subregion of the department were used, where the supervised learning models, Random Forest (RF), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and Naive Bayes (NB) were evaluated. **Results.** The results were analyzed using precision tests, kappa coefficient and confusion matrix. **Conclusion.** The RF algorithm demonstrated the best performance in estimating Boron levels, achieving 78% accuracy, outperforming SVM (75%), K-NN (69%) and NB (35%). **Implications.** These results allow for improved decision-making regarding fertilization and micronutrient management in the soil, in order to improve soil quality and therefore crop yield.

Key words: Prediction; algorithms; supervised learning; sustainable agriculture; soil nutrition; agroecosystems.

RESUMEN

Antecedentes. La región sur de Nariño en Colombia presenta una fuerte vocación agrícola. Sin embargo, se ha identificado que la tecnología asociada con la recomendación de fertilización especialmente en microelementos en el suelo es limitada, lo que genera conflictos en la producción. Esta situación demanda un enfoque diferente e innovador para abordar dichos desafíos. Es por ello que esta investigación utiliza la capacidad predictiva del enfoque de aprendizaje automático o Machine learning (ML). **Objetivo.** Explorar la aplicación de herramientas de ML para la predicción de los niveles de Boro en suelos Andisoles de Nariño e identificar el algoritmo más eficiente. **Metodología.** Se empleó 1,067 muestras de suelo recolectadas en diversos predios de cinco municipios en la subregión sur del departamento, donde se evaluó los modelos de aprendizaje supervisado, Random Forest (RF), K-Vecinos más Cercanos (K-NN), Máquina de Vectores de Soporte (SVM) y Naive Bayes (NB). **Resultados.** Los resultados fueron analizados mediante pruebas de precisión, coeficiente kappa y matriz de confusión. **Conclusión.** El algoritmo RF demostró el mejor rendimiento en la estimación de los niveles de Boro, logrando un 78% de precisión, superando a SVM (75%), K-NN (69%) y NB (35%). **Implicaciones.** Estos resultados permiten mejorar la toma de decisiones en cuanto a fertilización y gestión de micronutrientes en el suelo, con el fin de mejorar la calidad del suelo y por ende el rendimiento de los cultivos.

Palabras clave: Predicción; algoritmos; aprendizaje supervisado; agricultura sostenible; nutrición edáfica; agroecosistemas.

† Submitted April 24, 2024 – Accepted November 8, 2024. <http://doi.org/10.56369/tsaes.5595>



Copyright © the authors. Work licensed under a CC-BY 4.0 License. <https://creativecommons.org/licenses/by/4.0/>

ISSN: 1870-0462.

INTRODUCCIÓN

La nutrición vegetal abarca el suministro, absorción y asimilación de elementos químicos esenciales, en cantidades y momentos apropiados, para el crecimiento y desarrollo de las plantas (Hernández-Terrón *et al.*, 2021). Por tanto, comprender la dinámica de acumulación, y los requisitos y funciones de estos elementos es fundamental para lograr una producción sostenible (Zhu *et al.*, 2022).

En la categoría de nutrientes esenciales se distinguen los macroelementos necesarios en cantidades relativas mayores, y los microelementos que requieren menor cantidad por la planta. No obstante, un suministro nutricional que no contemple alguno de estos, interrumpe la asimilación fotosintética y el crecimiento de cualquier cultivo (Barbier, 2019).

El Boro, en calidad de micronutriente esencial interviene en diversos procesos metabólicos de las plantas, como la síntesis de componentes celulares, incluida la pectina; la carencia de este mineral puede resultar en paredes celulares frágiles y una menor capacidad para responder a lesiones mecánicas y amenazas patógenas (Acuña & Molina, 2018; Lagos, 2019). Además, desempeña un papel vital en la etapa productiva de la planta, al intervenir en los procesos de germinación del polen y el crecimiento del tubo polínico que inciden en la formación de frutos y semillas (Novoa *et al.*, 2018; Zhu *et al.*, 2022).

La insuficiencia de Boro es catalogada como la segunda deficiencia más relevante de microelementos en los cultivos, y plantea un desafío significativo para la agricultura (Barbier, 2019). A pesar de esta realidad, los agricultores se encuentran limitados en cuanto a estrategias de manejo, recurriendo a la aplicación de fertilizantes sintéticos para corregir las deficiencias en el suelo, estableciendo un panorama de sub-dosificación o sobredosificación (Camacho, 2020).

Los métodos de detección del micronutriente dependen del análisis químico y representan un costo alto especialmente para los sistemas de producción agrícola tradicional. Por ejemplo, Álvarez-Sánchez *et al.* (2020), establecieron que, en una muestra de cerca de mil fincas de la subregión sur de Nariño, menos del 5% de los sistemas evaluados realizaban análisis de suelo para establecer planes de fertilización de macroelementos, y ningún productor encuestado reportó el estudio de micronutrientes. Es por ello, que surge la necesidad de desarrollar nuevas metodologías para el diagnóstico de nutrientes del suelo y el manejo de datos agrícolas. En este contexto, las herramientas de Aprendizaje Automático o *Machine Learning* (ML) se han utilizado para descubrir patrones en diversos conjuntos de datos (Yepes *et al.*, 2022).

Cuando se establece un marco sólido, estas herramientas de ML pueden generar predicciones acertadas ofreciendo un método práctico y de bajo costo (Becerra-Rondón *et al.*, 2022; Mallah *et al.*, 2022; Zhu *et al.*, 2022). Recientemente, varios estudios han explorado la incorporación de esta estrategia en la toma de decisiones, buscando la sostenibilidad agrícola a través de la información (Meshram *et al.*, 2021).

Entre los algoritmos más utilizados en la predicción se destacan: Random Forest (RF), que es un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para realizar tareas de clasificación y regresión. Cada árbol se construye a partir de una muestra aleatoria del conjunto de datos, y en cada división se selecciona un subconjunto aleatorio de características, lo que ayuda a mejorar la precisión y a reducir el sobreajuste. La predicción final se obtiene mediante el voto mayoritario en el caso de clasificación o el promedio en el caso de regresión de las predicciones de todos los árboles del bosque (Rajković *et al.*, 2021; Zolfaghari *et al.*, 2023), Vecinos más Cercanos (K-NN) es un método de aprendizaje supervisado utilizado para clasificación y regresión. Se basa en la premisa de que datos similares están cercanos en el espacio de características. Para clasificar un nuevo punto, el algoritmo identifica los *k* vecinos más cercanos en el conjunto de entrenamiento, generalmente utilizando una medida de distancia, y asigna la clase más frecuente entre estos vecinos o calcula el promedio de sus valores (Abbas *et al.*, 2020; Bouasria *et al.*, 2020), Máquinas de Vectores de Soporte (SVM) que es un algoritmo utilizado principalmente para clasificación y, en menor medida, para regresión. Se basa en la identificación de un hiperplano óptimo que separa las clases en un espacio de características de alta dimensión, maximizando el margen entre los vectores de soporte (puntos de datos más cercanos al hiperplano). SVM puede utilizar núcleos (kernels) para transformar datos no linealmente separables en un espacio donde se pueda encontrar una separación lineal (Zhang *et al.*, 2019; Ismaili *et al.*, 2023), y Naive Bayes (NB), que es un conjunto de algoritmos de clasificación probabilística basados en el teorema de Bayes, que asume la independencia condicional entre las características de entrada. Este enfoque permite calcular la probabilidad de una clase dada un conjunto de características, utilizando la información sobre la frecuencia de las clases y las características en los datos (Ramírez, 2020). Cada algoritmo presenta características distintivas y una capacidad de adaptación a diferentes tipos de problemas y conjuntos de datos, por lo cual, cada situación problema requiere una evaluación específica (Álvarez *et al.*, 2023).

Considerando lo anteriormente mencionado, esta investigación buscó evaluar diferentes algoritmos de

aprendizaje automático para predecir el contenido de Boro en suelos agrícolas de la subregión sur del departamento de Nariño.

METODOLOGÍA

Los datos provienen de un estudio realizado en la subregión sur del departamento de Nariño por Álvarez *et al.* (2019), en donde se realizó un muestreo aleatorio a 1,256 fincas en los municipios de Ipiales (683), Córdoba (230), Cuaspué (223), Potosí (60) e Iles (60). Los predios evaluados se situaron entre los 1,829 y 2,847 msnm, con una distribución espacial que varió entre -77.8 y -77.4 grados de longitud y 0.9 a 1.6 grados de latitud.

Con el objetivo de caracterizar el orden de suelo del área de estudio, se utilizó la capa de Suelos de Nariño a escala 1:100,000 del Instituto Geográfico Agustín Codazzi – IGAC provista por el geovisor de Colombia en mapas (<https://www.colombiainmapas.gov.co/>). Esta capa de uso del suelo se recortó con la unión de los polígonos de los municipios del área de interés y posteriormente se dio formato a la simbología, teniendo en cuenta el perfil modal. Las ubicaciones tomadas con GPS de los lugares donde se realizaron los puntos de muestreo se cruzaron con el mapa general de suelos del departamento. A partir de esta información con ayuda de la memoria técnica, se determinó las características de los perfiles modales de suelos y su clasificación (Soil Survey Staff, 1999).

Una vez confirmado el orden Andisol en cada finca se obtuvo una muestra representativa de suelo en los primeros 20 cm de profundidad, tomando un promedio de 20 puntos por hectárea (IGAC, 2017), estos se mezclaron para formar una sola muestra compuesta debidamente rotulada. En laboratorio se determinó pH mediante método potenciométrico, conductividad eléctrica mediante método electrométrico, Capacidad de Intercambio Catiónico Efectiva (CICE), textura mediante método de Bouyucos, Materia orgánica (MO) mediante método de Walkley- Black, Fosforo (P) mediante Bray II, Nitrógeno total mediante método Kjeldahl, Aluminio (Al) mediante método KCl, Azufre (S) mediante método de Turbidimetría, las bases intercambiable Ca, Mg y K mediante método de acetato de amonio y el elemento menor Boro mediante el método de Azometrina.

Interfaz de trabajo

La implementación de los procedimientos estadísticos se llevó a cabo mediante el uso del lenguaje de programación R, versión 4.1.2 (R Development Core Team, 2013), empleando las librerías especializadas, MASS, psych, caret, randomForest, rpart y rpart.plot.

Exploración estadística de variables

Para cada una de las variables se evaluó el comportamiento de los datos realizando un análisis descriptivo, es decir, mínimo, máximo, media y coeficiente de variación. Esta fase permitió detectar valores atípicos y aquellos que no correspondieron al orden Andisol, los cuales fueron eliminados del conjunto de datos quedando finalmente con un consolidado de 1,013 análisis de suelo. Con el objetivo de determinar el grado de asociación entre variables y evitar la multicolinealidad, se aplicó una prueba de correlación de Pearson, optando por seleccionar una de las variables cuando el coeficiente presento un valor superior a 0.6.

Conformación de grupos a predecir

En el proceso de formación de los grupos se empleó la variable original de concentración de Boro medida en ppm. La asignación de cada muestra a un grupo se realizó considerando los cuartiles de esta variable, lo que resultó en el arreglo de cuatro grupos que representaron el 25% de los datos originales. Esta estrategia garantizó que cada grupo contemple una proporción relativamente homogénea de la variabilidad en los niveles de Boro, facilitando la interpretación y evaluación de los resultados de la predicción.

Obtención del set de trabajo

Los datos se dividieron aleatoriamente en tres conjuntos: entrenamiento (70% de los datos), prueba (20% de los datos) y testeo (10% de los datos). El conjunto de entrenamiento se destinó a preparar los modelos de aprendizaje automático implementando un proceso de validación cruzada, con el fin de evaluar el rendimiento del modelo de manera robusta a partir de diez subconjuntos aleatorios. El conjunto de prueba se empleó para evaluar el rendimiento posterior al entrenamiento. Por su parte, los datos de testeo se utilizaron para generar una matriz de confusión, proporcionando así una evaluación detallada del rendimiento de los modelos. Se empleó la librería Care, que emplea técnicas de normalización min-max para escalar las variables predictoras. Este proceso se realizó con el fin de transformar los valores de las características en un rango definido.

Modelos de aprendizaje a evaluar

Se evaluaron cuatro modelos de aprendizaje automático en el contexto de predicción supervisada, los cuales fueron probados para asignar etiquetas o categorías a conjuntos de datos fundamentándose en patrones y características identificables. Entre los modelos evaluados están Random Forest (RF), K-Vecinos más Cercanos (K-NN), Máquina de Vectores

de Soporte (SVM) y Naive Bayes (NB).

Para cada algoritmo se realizaron ajustes independientes en algunos parámetros que influyen en la precisión de este. En cuanto a RF, los parámetros de ajuste fueron mtry que osciló entre 5 a 6, y el tamaño mínimo de nodo que osciló entre 2 y 4. Para SVM la variación en el parámetro C estuvo entre 50 a 800, y entre 0.5 a 1 para sigma. Para el algoritmo K-NN el valor de K varió entre 1 a 50, y para NB los parámetros de ajuste fueron fL y Adjust que se ubicaron entre 0 y 1 en ambos casos.

Métricas de evaluación

El desempeño de los modelos se basó en tres métricas de evaluación, para las dos primeras se determinó el número de verdaderos positivos (TP) y verdaderos negativos (TN), que son el número de observaciones identificados correctamente por los algoritmos. Falsos positivos (FP), indicando la cantidad de observaciones identificadas incorrectamente por los algoritmos y falsos negativos (FN), que indica el número de observaciones no reconocidas (Ismaili *et al.*, 2023).

La precisión, corresponde a una medida de las instancias predichas correctamente y está determinada

por la ecuación:

$$Precisión = \frac{(TP+TN)}{((TP+TN+FP+FN))} \quad (\text{Ecuación 1})$$

El coeficiente Kappa, mide el nivel de concordancia entre las clases predichas y reales, empleando la ecuación 2:

$$Kappa = \frac{(TP+TN)}{((TP+TN+FP+FN))} \quad (\text{Ecuación 2})$$

Finalmente, se generó una matriz de confusión con el set de testeo; los resultados se almacenaron en una tabla que resume los valores de recuento desglosados por cada clase, en las columnas se presentan los valores reales y en las filas los valores predichos.

RESULTADOS Y DISCUSIÓN

Exploración estadística de variables

Basándonos en los perfiles modales de suelo de las 1,256 muestras iniciales, se excluyeron 189 fincas que no cumplían con las características del orden Andisol, resultando en un conjunto de datos final de 1,067 muestras de suelo como se muestra en la Figura 1.

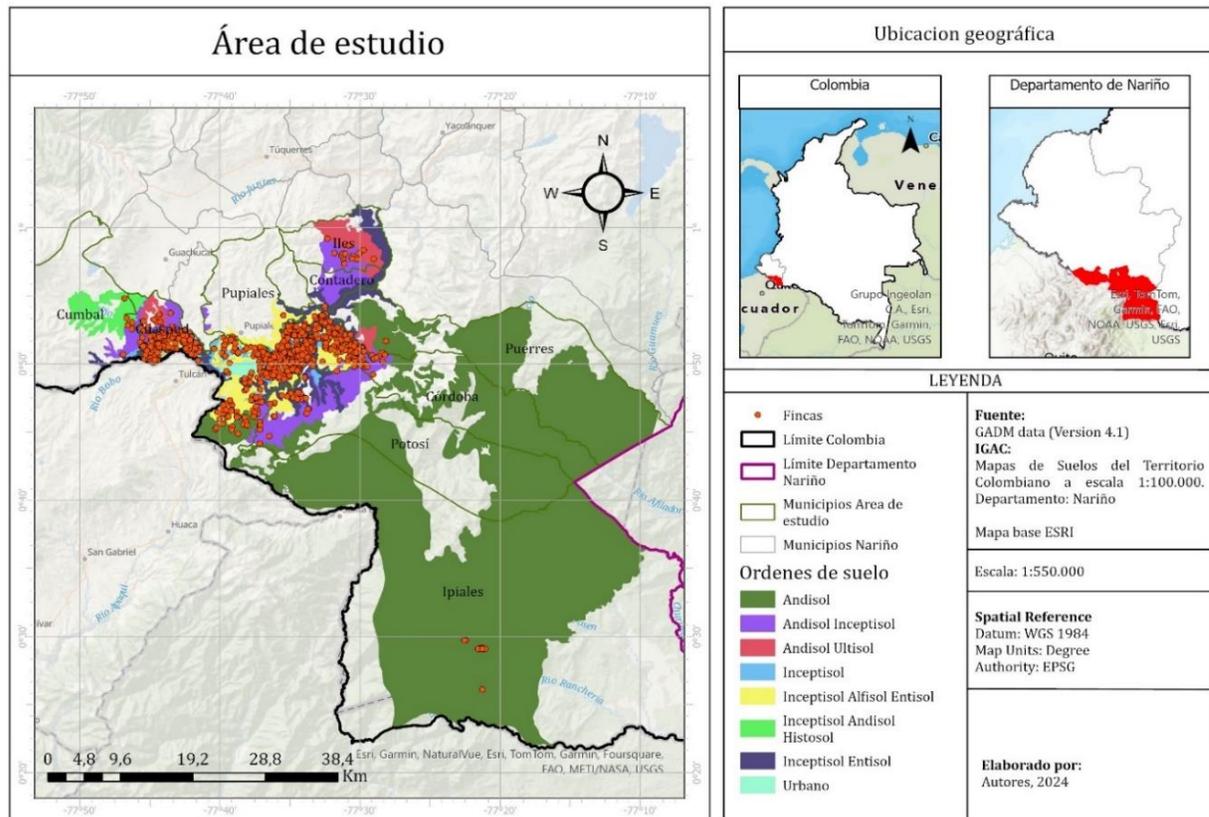


Figura 1. Distribución espacial de puntos de muestreo y caracterización de órdenes de suelo en la subregión sur del departamento de Nariño (Colombia).

En la zona evaluada se evidenció una acidez desde fuertemente ácida a modernamente ácida, con un valor medio de pH de 5.05 (Tabla 1), corroborando lo encontrado por Culchac *et al.* (2021), y Alzate *et al.* (2022), quienes reportan valores de 4 a 5.7; este rango de pH (4 a 5.7) fue similar al indicado por Castro-Rincón *et al.* (2019), con 11.9 cmol/kg, pero contrastante con reporte de Lagos y Huertas. (2019), quienes encontraron valores de 41.77 cmol/kg. La variabilidad detectada podría vincularse con la aplicación de enmiendas orgánicas en ciertas fincas antes del período de muestreo, lo que motivó la exclusión de datos atípicos.

Por su parte, los contenidos de Aluminio son moderados, mostrando una media de 0.47 cmol/kg (Tabla 1). En la investigación de Medina *et al.* (2019), se reportaron valores promedio inferiores a 2 cmol/kg, resaltando que la acidez del suelo es una de las principales características que disminuye la disponibilidad de nutrientes para las plantas la región.

Asociado a lo anterior, Camacho *et al.* (2020), indican que la acidez también incide en los cambios de pH e intensifica la proporción de Fósforo fijado en la matriz del suelo, lo que concuerda con los bajos niveles de este elemento encontrado en esta investigación, en el orden de 3.01 a 12.5 ppm con un valor medio de 7.06 ppm (Tabla 1), dicho valor fue cercano al valor promedio de 8.3 ppm reportados por Castro-Rincón *et al.* (2019).

El contenido de Materia Orgánica con un promedio de 6.52% se clasificó como medio (Tabla 1), resultado que posiblemente se deba a que los predios evaluados se encontraban en regiones altas, en donde se presentan procesos de menor velocidad de mineralización, por tanto, la acumulación de restos orgánicos es constante (Lagos y Huertas, 2019). Asociado a este hecho, el análisis fisicoquímico reveló que el promedio de contenido de Nitrógeno en la subregión sur de Nariño fue de 0.21%, valor cercano al reportado por Álvarez *et al.* (2021), quienes encontraron un contenido de 0.19% en la misma subregión, y el contenido de Azufre presentó una media de 9.45 mg/kg similar a los resultados reportados por Rodríguez-Rodríguez *et al.* (2022).

El contenido de las bases intercambiables Calcio, Magnesio y Potasio fue clasificado como medio (Tabla 1). Las relaciones entre estos elementos también fueron evaluadas, mostrando una proporción Ca/Mg de 2.7, Ca/K de 6.8 y Ca+Mg/K de 25. Estos valores proporcionan una perspectiva de la disponibilidad relativa de estos cationes en el suelo; este hallazgo se alinea con el resultado obtenido para Capacidad de Intercambio Catiónico Efectiva con un valor medio de 9.7 cmol/kg sugiriendo una adecuada capacidad del suelo para retener y liberar cationes.

Por su parte, las variables arcilla, arena y limo presentaron valores que se encuentran dentro de los rangos reportados por Volverás-Mambuscay *et al.* (2020); para arcilla se reporta valores cercanos al 33%, para arena 55.5% y para limo entre 19 y 32.8%; los valores que evidencian distinciones particulares en la muestra podrían atribuirse a la heterogeneidad del suelo (manejo, topografía, altitud, génesis del suelo), dado que las características fisicoquímicas y biológicas experimentan cambios locales incluso a distancias reducidas (Marcillo- Paguay *et al.*, 2022).

De igual forma, en relación con la matriz de datos se llevó a cabo un proceso de selección de variables predictoras con el propósito de asegurar la ausencia de correlación significativa entre ellas. Es decir, que la información proporcionada por una variable no esté redundante con la información suministrada por otra (Kebonye *et al.*, 2023). Este procedimiento para Ennaji *et al.* (2023), y El-Sayed *et al.* (2023), se erige como un paso esencial ya que no solamente puede potenciar la eficiencia del modelo, sino que, además, ayuda a mitigar el riesgo de sobreajuste.

La Tabla 2 revela una asociación fuerte entre las partículas que componen la textura del suelo ($L \times A = -0.96$, $Ar \times A = -0.50$ y $Ar \times L = 0.43$). Lagos y Huertas (2019) y Volverás-Mambuscay *et al.* (2020), consideran que las arenas son partículas inertes desde el punto de vista químico, los limos están en un punto intermedio en cuanto a actividad química y propiedades físicas, mientras que las arcillas se comportan como coloides con carga eléctrica con capacidad para intercambio iónico. Por esto, se tomó la decisión de incorporar solamente a esta última variable en el entrenamiento.

Tabla 1. Variables fisicoquímicas de suelos Andisoles de la Subregión sur de Nariño (n=1,013).

| Estadístico | pH | MO (%) | N | P | S (ppm) | B | K | Ca | Mg (cmol/Kg) | Al | CICE | Ar | Ac (%) | Li |
|-------------|------|--------|-------|-------|---------|-------|-------|-------|--------------|-------|-------|-------|--------|-------|
| Media | 5.05 | 6.52 | 0.21 | 7.06 | 9.45 | 0.35 | 0.36 | 6.84 | 2.47 | 0.47 | 9.70 | 44.29 | 31.33 | 24.37 |
| Mínimo | 3.96 | 4.21 | 0.14 | 3.01 | 4.10 | 0.27 | 0.23 | 2.00 | 0.60 | 0.01 | 0.01 | 6.00 | 21.00 | 1.00 |
| Máximo | 6.17 | 8.91 | 0.29 | 12.15 | 13.80 | 0.49 | 0.51 | 12.9 | 4.50 | 0.99 | 17.0 | 70.00 | 42.00 | 67.00 |
| CV (%) | 8.10 | 10.95 | 11.28 | 25.34 | 34.45 | 11.60 | 14.25 | 28.60 | 24.15 | 55.93 | 24.27 | 34.45 | 14.35 | 55.85 |

Nota: *Ar: Arena; Ac: Arcillas; Li: Limos.

Tabla 2. Análisis de correlación de variables fisicoquímicas del suelo subregión sur de Nariño (n=1,013).

| Variables | A | Ar | L | pH | MO | N | P | S | Ca | Mg | K | CICE | Al | B |
|---------------------|--------------|-------------|-------|--------------|-------------|------|-------|-------|-------------|-------------|-------|-------|-------|---|
| Arena (A) | 1 | | | | | | | | | | | | | |
| Arcilla (Ar) | -0.50 | 1 | | | | | | | | | | | | |
| Limo (L) | -0.96 | 0.43 | 1 | | | | | | | | | | | |
| pH | -0.01 | 0.06 | -0.01 | 1 | | | | | | | | | | |
| MO | 0.09 | -0.07 | -0.08 | -0.11 | 1 | | | | | | | | | |
| N | 0.09 | -0.07 | -0.08 | -0.11 | 0.99 | 1 | | | | | | | | |
| P | -0.02 | 0.04 | 0.01 | -0.07 | 0.11 | 0.11 | 1 | | | | | | | |
| S | 0.03 | -0.02 | -0.03 | -0.04 | 0.04 | 0.05 | 0 | 1 | | | | | | |
| Ca | 0.03 | 0.04 | -0.05 | 0.1 | 0.01 | 0.02 | -0.04 | 0 | 1 | | | | | |
| Mg | 0.02 | 0.05 | -0.04 | 0.07 | 0 | 0 | -0.06 | -0.02 | 0.51 | 1 | | | | |
| K | 0.18 | -0.06 | -0.18 | -0.02 | 0.06 | 0.05 | -0.01 | 0.02 | -0.05 | -0.02 | 1 | | | |
| CICE | 0.04 | 0.04 | -0.06 | 0.1 | 0.02 | 0.02 | -0.06 | -0.01 | 0.97 | 0.68 | -0.03 | 1 | | |
| Al | -0.01 | -0.03 | 0.02 | -0.88 | 0.13 | 0.13 | 0.08 | 0.04 | -0.08 | -0.07 | 0.05 | -0.09 | 1 | |
| B | 0.09 | 0 | -0.1 | 0.06 | 0.03 | 0.03 | 0.02 | -0.04 | -0.03 | -0.01 | 0.26 | -0.03 | -0.08 | 1 |

La correlación significativa entre Materia Orgánica y Nitrógeno también descrita en la Tabla 2, se explica por el hecho de que el material orgánico después de pasar por procesos de mineralización se convierte en la principal fuente de N para el suelo (Apráez *et al.*, 2019; Marcillo-Paguay *et al.*, 2022). Además, según Mason-Jones *et al.* (2018), la MO presenta mayor sensibilidad a modificaciones ambientales, hecho que se tuvo en cuenta para su selección.

Las variables CICE, Ca⁺, y Mg exhibieron una clara asociación, ya que el ion Ca⁺ junto con el Mg son componentes determinantes del CICE; por consiguiente, se decidió continuar el análisis enfocándose en la variable compuesta. Esta elección se fundamenta en la capacidad de la CICE para proporcionar información acerca de la habilidad del suelo para retener cationes, además de ser una variable que informa sobre la disponibilidad de estos (Marcillo-Paguay *et al.*, 2022).

Por último, la concentración de iones de aluminio (Al⁺) está estrechamente vinculada a la acidez del suelo, ya que los iones de Al⁺ intercambiables forman compuestos cuyas reacciones generan iones H⁺ que contribuyen al aumento de la acidez (Camacho *et al.*, 2020). Asimismo, se reconoce que el pH del suelo desempeña un papel determinante en la disponibilidad de los elementos esenciales para el desarrollo de las plantas, siendo para fines de este estudio de mayor valor analítico como lo sugiere Medina *et al.* (2019).

Basados en las asociaciones anteriores se procedió a seleccionar ocho variables predictoras Ar, pH, MO, P, S, K, CICE y Mg, que fueron empleadas para evaluar los cuatro modelos de ML. De esta manera se buscó abarcar factores que pueden integrar al proceso de cómputo relaciones e interdependencia con la concentración de Boro en el suelo, para proporcionar una herramienta de predicción eficiente.

Conformación de grupos a predecir

La distribución de la variable Boro en la muestra analizada reveló que el Grupo 1 (G1), compuesto por los valores situados en el primer cuartil (q1), exhibe una concentración del micronutriente inferior a 0.32 ppm. Por su parte, el Grupo 2 (G2) que abarcó los valores desde el primer cuartil hasta la mediana, muestra concentraciones entre 0.32 y 0.36 ppm (Figura 2). Estos primeros grupos mostraron niveles deficientes de Boro en el suelo, lo que señala la necesidad de atención especial por parte del agricultor en términos de manejo (Choi *et al.*, 2015).

El Grupo 3 (G3) abarcó los valores con concentraciones desde la mediana hasta el tercer cuartil (q3), oscilando entre 0.36 y 0.39 ppm y finalmente, el Grupo 4 (G4) incluye los valores con concentraciones superiores al tercer cuartil, es decir, mayores a 0.39 ppm (Figura 2). La concentración de Boro en este caso es considerada como intermedia según el rango establecido por Choi *et al.* (2015).

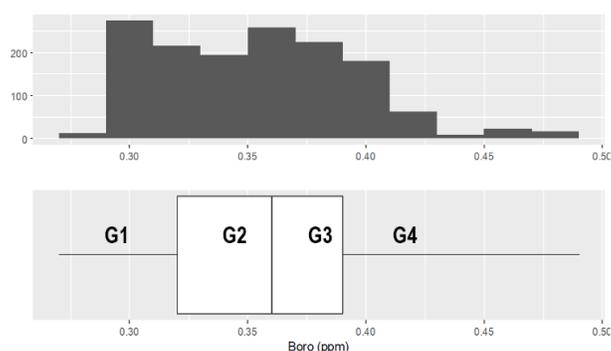


Figura 2. Distribución de la variable Boro y conformación de grupos a predecir (n=1,013).

Es relevante destacar que, en los 1,067 análisis de suelos evaluados, el valor máximo reportado fue de 0.49 ppm de Boro, por lo cual, se recomienda una investigación más exhaustiva al respecto, dado que esta concentración podría evolucionar como una posible limitante para la nutrición de los cultivos. Además, se debe considerar que esta situación podría agravarse en terrenos con topografía quebrada, propios de la región sur de Nariño como lo detalla Álvarez *et al.* (2019), dada la reducida concentración de este micronutriente, lo cual podría estar vinculado con prácticas inadecuadas de manejo del suelo en la región (Zhu *et al.*, 2022).

Evaluación de los modelos de predicción

En la Tabla 3, se muestra el rendimiento de los algoritmos de aprendizaje automático evaluados, en donde tres de ellos presentaron métricas de rendimiento de 0.7 y 0.54 en precisión y coeficiente kappa respectivamente. Estos valores indican que las estimaciones respecto al valor real de predicción son confiables, evidenciando un grado de concordancia aceptable del procedimiento (Álvarez *et al.*, 2019).

Por otra parte, en este estudio se pudo establecer que en promedio se presentó una variación del 0.1 en precisión y kappa al contrastar los resultados de entrenamiento y prueba; esta fluctuación en las métricas de rendimiento establece que los modelos no están sobreajustados a los datos de entrenamiento, lo cual es fundamental para asegurar su capacidad de

generalización de conjuntos de datos no conocidos.

El algoritmo RF tuvo las métricas de mayor eficiencia (Tabla 3), demostrando su capacidad para adecuarse a diferentes situaciones de predicción, puesto que, su configuración se basa en la generación de resultados a partir de múltiples árboles de decisión tomando datos seleccionados aleatoriamente con el fin de reducir el sobreajuste. En este sentido, un conjunto de datos relativamente pequeño puede analizarse favorablemente debido a la fuerte regularización (Becerra-Rondón *et al.*, 2022). Así mismo, diferentes investigadores afirman que el modelo generado por RF ofrece una relación señal-ruido baja, lo cual ha extendido este algoritmo por diferentes ámbitos de investigación (Yepes *et al.*, 2022).

Zolfaghari *et al.* (2023), llevaron a cabo una evaluación de la variabilidad espacial de parámetros como N, K, Mo, pH, Ca, A, L, Ar y Dap utilizando información fisicoquímica del suelo, datos climáticos e imágenes de satélite, empleando modelos de Redes Neuronales Artificiales, árboles de regresión cubista, RF y K-NN. Los resultados indicaron que, de estos modelos, RF destacó como el más efectivo para estimar N, K, C, pH y Dap en el suelo. Esta conclusión también se alinea con otros estudios exitosos que utilizaron RF para predecir variables del suelo, como los trabajos de Mallah *et al.* (2022), Zhu *et al.* (2022) y Becerra-Rondón *et al.* (2022), donde, si bien no siempre RF se identifica como el modelo más eficiente, en cada caso exhibe un rendimiento destacado.

Tabla 3. Configuración de los algoritmos para predecir Boro en suelos Andisoles.

| Algoritmo | Parámetro 1 | Parámetro 2 | Entrenamiento | | Prueba | |
|---------------|----------------|----------------------------|---------------|-------------|-------------|-------------|
| | | | Precisión | Kappa | Precisión | Kappa |
| Random Forest | mtry=5 | Tamaño.min.nodo = 2 | 0.63 | 0.50 | 0.64 | 0.52 |
| | mtry= 5 | Tamaño.min.nodo = 3 | 0.62 | 0.50 | 0.63 | 0.61 |
| | mtry=5 | Tamaño.min.nodo = 4 | 0.64 | 0.50 | 0.64 | 0.52 |
| | mtry=6 | Tamaño.min.nodo = 2 | 0.63 | 0.51 | 0.63 | 0.51 |
| | mtry= 6 | Tamaño.min.nodo = 3 | 0.79 | 0.67 | 0.78 | 0.65 |
| | mtry=6 | Tamaño.min.nodo = 4 | 0.63 | 0.50 | 0.64 | 0.52 |
| SVM | C=50 | Sigma= 1 | 0.75 | 0.59 | 0.75 | 0.59 |
| | C=100 | Sigma= 1 | 0.65 | 0.52 | 0.66 | 0.55 |
| | C=700 | Sigma= 1 | 0.65 | 0.52 | 0.66 | 0.55 |
| | C=800 | Sigma= 0,5 | 0.64 | 0.52 | 0.66 | 0.55 |
| KNN | | K=1 | 0.70 | 0.54 | 0.69 | 0.55 |
| | | K=2 | 0.36 | 0.34 | 0.36 | 0.30 |
| | | K=5 | 0.35 | 0.13 | 0.57 | 0.30 |
| | | K=10 | 0.34 | 0.10 | 0.57 | 0.31 |
| | | K=20 | 0.34 | 0.10 | 0.30 | 0.20 |
| | | K=50 | 0.32 | 0.08 | 0.30 | 0.20 |
| NB | fL = 0 | Adjust= 0 | 0.33 | 0.09 | 0.35 | 0.10 |
| | fL = 0 | Adjust= 0,5 | 0.33 | 0.08 | 0.34 | 0.09 |
| | fL = 0 | Adjust= 1,0 | 0.33 | 0.07 | 0.33 | 0.06 |
| | fL = 1 | Adjust= 0 | 0.33 | 0.08 | 0.33 | 0.09 |
| | fL = 1 | Adjust= 0,5 | 0.33 | 0.08 | 0.33 | 0.06 |
| | fL = 1 | Adjust= 1,0 | 0.33 | 0.07 | 0.33 | 0.06 |

Los parámetros ajustados del algoritmo RF. En este caso, se fijó el número de variables en la primera división conocido como "mtry", en 6. Esto implica que, en cada nodo del árbol el algoritmo considera aleatoriamente 6 características para realizar una división. Además, se estableció un tamaño mínimo de nodo de 3, indicando que un nodo terminal debe contener al menos 3 observaciones. Esta configuración condujo a la obtención de la precisión y el valor kappa más elevados alcanzando valores de 0.78 y 0.65 respectivamente (Tabla 3).

El rendimiento destacado del modelo RF se traduce en beneficios significativos en la predicción del Boro, determinando el nivel del nutriente, lo cual es útil para generar estrategias de nutrición y calculo adecuado de aportes de Boro cuando sea necesario. La capacidad del algoritmo para proporcionar información detallada sobre la disponibilidad del micronutriente, sería importante en la producción de cultivos bajo entornos similares al estudio.

Además, para validar los resultados de generalización de cada algoritmo se utilizó la matriz de confusión (Figura 3). En ella, se observa que RF registró el menor número de casos incorrectos para los grupos G1 y G2 evidenciado por el mayor número de observaciones sobre la diagonal principal. En contraste, los grupos G3 y G4 presentaron una proporción más alta de registros clasificados erróneamente, sin diferencias notables entre estas dos clases en términos de verdaderos positivos y falsos negativos.

| | | RF | | | | SVM | | | |
|----|--|----|----|----|----|-----|----|----|----|
| | | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |
| G1 | | 44 | 9 | 0 | 0 | 40 | 9 | 4 | 0 |
| G2 | | 4 | 42 | 7 | 0 | 5 | 42 | 6 | 0 |
| G3 | | 0 | 8 | 40 | 5 | 0 | 11 | 38 | 4 |
| G4 | | 0 | 2 | 10 | 41 | 0 | 3 | 11 | 39 |
| | | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |

| | | KNN | | | | NB | | | |
|----|--|-----|----|----|----|----|----|----|----|
| | | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |
| G1 | | 44 | 7 | 2 | 0 | 15 | 17 | 12 | 9 |
| G2 | | 4 | 43 | 6 | 0 | 17 | 18 | 13 | 5 |
| G3 | | 2 | 8 | 37 | 6 | 7 | 11 | 23 | 12 |
| G4 | | 1 | 3 | 6 | 43 | 8 | 10 | 16 | 19 |
| | | G1 | G2 | G3 | G4 | G1 | G2 | G3 | G4 |

Figura 3. Matriz de confusión para validación de los datos. Grupo de testeo (n=212).

El segundo algoritmo que mostró un rendimiento importante fue SVM, el cual se fundamenta en la generación de hiperplanos entre los diversos grupos de datos mediante vectores extremos (Ismaili *et al.*,

2023). Luego, calcula un límite de margen máximo que conduce a una partición de los grupos, generando reglas de predicción tanto para datos que son separables linealmente como para aquellos que no lo son (Aworka *et al.*, 2022).

El menor poder predictivo de SVM, podría atribuirse a una posible falta de identificación adecuada de los parámetros de aprendizaje. En este sentido, estudios previos sugieren que antes de emplear SVM, es recomendable utilizar conjuntos de datos amplios, que abarquen diversas regiones o variaciones para determinar los valores óptimos de sus parámetros (Bayat *et al.*, 2020). Por ejemplo, Abraham *et al.* (2020), al evaluar SVM en la clasificación de suelos en grupos hidrológicos encontraron una precisión promedio del 72%, utilizando las variables predictoras A, L, Ar y conductividad hidráulica saturada, con un número de observaciones tres veces mayor al empleado en este estudio. Sin embargo, es relevante indicar que estos autores optaron por algoritmos como RF, debido a la rapidez en la parametrización y la interpretación de los resultados.

Así mismo, en la Tabla 3 se observa que para el algoritmo SVM seleccionando un valor C de 50 y un valor sigma de 1, se obtiene una precisión del 0.75 y un kappa de 0.59. Lo anterior, indica que el valor de sigma fue favorable para obtener un límite de decisión más flexible, mientras que el parámetro C influye en un sesgo alto y una varianza baja. Valores más altos de sigma pueden resultar en límites de decisión más suaves y adaptables, lo que puede ser beneficioso en situaciones donde los datos presentan una mayor variabilidad o complejidad. Por otro lado, valores más bajos de sigma pueden llevar a límites de decisión más rígidos y simplificados, lo que puede ser útil en casos donde se busca una separación más clara entre las clases.

El tercer algoritmo de acuerdo al rendimiento fue K-NN, un algoritmo no paramétrico que basa su funcionamiento en el principio de similitud y proximidad de los datos (Bouasria *et al.*, 2020). K-NN asigna un valor a la nueva ubicación que está más cerca de los K puntos de entrenamiento dentro del espacio auxiliar multivariante, y utiliza la similitud de características para predecir el clúster en el que caerá el nuevo punto (Ramírez, 2020; Zolfaghari *et al.*, 2023). Posiblemente la razón por la que K-NN no mostró un comportamiento favorable en este estudio, se deba a que únicamente utiliza la distancia para clasificar, e ignora la distribución espacial de los datos (Zolfaghari *et al.*, 2023). Al no tener en cuenta dicha distribución y debido a que los datos varían significativamente de un sitio a otro, se limitan ciertas características a extraer de los datos de entrada del modelo y, por tanto, se tienen dificultades para generalizar correctamente como lo confirman Pant *et*

al. (2019).

Además, el valor de K para clasificar cada nuevo dato es fijo y único, cuando el número de vecinos necesarios debería de ser distinto en función de si el nuevo dato se encuentra en el centro de una clase de entrenamiento, o si se halla en la superposición de varias clases (Bouasria *et al.*, 2020). En el estudio de Mohd *et al.* (2022), se sugiere evaluar un rango de entre 1 a 10 valores, con el fin de obtener el número óptimo de vecinos para cada grupo dentro de los K puntos de entrenamiento, explorando con cada vecino un espectro más amplio de similitud de características.

En este estudio conforme aumentó el número de vecinos la precisión tiende a disminuir (Tabla 3), lo cual es corroborado por Galupino y Dungca. (2022), quienes afirman que al aumentar el parámetro K los resultados se estabilizan al tiempo que aumenta la inexactitud. Sin embargo, con un valor de K=1, se vaticina una alta probabilidad de sobreajuste para los datos de entrenamiento lo que podría traducirse en una posible falta de generalización cuando se utilice en la tarea asignada, por ello se aprecia una disminución de la precisión del algoritmo en la fase de prueba.

Finalmente, NB obtuvo el rendimiento más bajo (Tabla 3). Este algoritmo se basa en construir una hipótesis para el conjunto de clases de salida en función de los atributos de entrada, mediante la generación de una distribución normal entre clases a través de una estimación de máxima verosimilitud (Mohd *et al.*, 2022; Ramírez, 2020). El bajo rendimiento de NB puede deberse a que el modelo es sensible a la cantidad de datos de entrenamiento, pues necesita suficientes ejemplos para establecer las probabilidades de ocurrencia de las diferentes características (Kaushik *et al.*, 2022). Si se tiene pocos datos el rendimiento de NB puede verse afectado, además, asume independencia condicional entre las características, lo que puede afectar su rendimiento en situaciones donde esta suposición no se cumple (Chen *et al.*, 2020).

Malik *et al.* (2021), realizaron un análisis comparativo de las propiedades del suelo para predecir la fertilidad y el rendimiento de los cultivos, utilizando algoritmos de aprendizaje automático como K-NN, NB y RF, encontrando que NB predijo con menor precisión el rango óptimo de pH, humedad y temperatura en muestras de suelo. En este estudio, el modelo obtuvo en la etapa de entrenamiento una precisión de 0.34 y un Kappa de 0.09, en donde, los resultados en el conjunto de prueba muestran una precisión del 0.35, ligeramente superior a la del conjunto de entrenamiento y un Kappa de 0.10 (Tabla 3). Igualmente, la baja precisión observada en la matriz de confusión (Figura 3), podría indicar que el modelo tiene dificultades para ajustarse a los datos de

validación generando una deficiente generalización.

Por tanto, el uso de KNN y NB para esta investigación incrementaron la posibilidad para que en el proceso de predicción se generen falsos positivos o falsos negativos, con consecuencias para la toma de decisiones ya que se asignarían valores erróneos al predecir los niveles de Boro, siendo esto crítico en suelos deficientes en este elemento, lo que finalmente podría repercutir en la productividad de los cultivos.

Se debe considerar, que el aumento del número de sitios de muestreo y la inclusión de variables cruciales para la cuantificación de los contenidos de Boro en el suelo, podrían ser clave para mejorar la precisión de los modelos. Al respecto Barbier (2019), señala que la disponibilidad de Boro para las plantas está fuertemente influenciada por características fisicoquímicas del suelo, como el tipo de suelo, la textura, los ciclos de humedad/sequía, la temperatura, el contenido de materia orgánica, arcillas, pH, el material parental local, la descomposición de minerales entre otros. Estas características se consideran insumos relevantes en nuevos proyectos de ML y pueden potenciar la capacidad predictiva de los modelos evaluados.

El destacado rendimiento de RF, lo posiciona como el algoritmo recomendado para seguir avanzando en el desarrollo de predicción de los niveles de Boro en suelos agrícolas. La aplicación de análisis de datos en la agricultura proporciona a los productores la capacidad de tomar decisiones más informadas, mejorando así la gestión del recurso suelo, aumentando la eficiencia de los cultivos y reduciendo el impacto ambiental (Meshram *et al.*, 2021).

Este estudio presenta resultados preliminares, que, aunque son prometedores requieren refinamiento, ajuste y bases de datos más robustas para llevar a cabo futuras investigaciones que consoliden aún más la eficacia de esta metodología en la mejora de la productividad agrícola.

CONCLUSIONES

El empleo de modelos de ML en el análisis de datos, representa una herramienta innovadora y una alternativa a los métodos tradicionales de evaluación fisicoquímica del suelo. Esto posibilita la predicción de minerales secundarios como el Boro manera rápida, precisa, eficiente y económica, mejorando así la gestión de decisiones para el agricultor y el asistente técnico.

Los análisis de suelos evaluados revelan un máximo de 0.49 ppm de Boro, valor que denota posibles limitaciones nutricionales en los suelos de la región sur de Nariño. Estos resultados subrayan la necesidad de

investigaciones adicionales y la implementación de prácticas de manejo del suelo adecuadas, pues la gestión eficaz es requerida para preservar la salud del suelo y asegurar la sostenibilidad agrícola a largo plazo.

El algoritmo RF demostró ser efectivo en la predicción de la disponibilidad de Boro en suelos Andisoles. Su capacidad para manejar la complejidad y variabilidad de los datos, junto con su robustez en la identificación de patrones, lo posiciona como una herramienta valiosa para la evaluación y gestión de nutrientes en estos tipos de suelo, permitiendo así mejorar la toma de decisiones en la agricultura y la conservación del medio ambiente.

Funding. No specific funding was used to support this research.

Conflict of Interest. The authors have no conflicts of interest to declare.

Compliance with ethical standards. This paper is an original contribution and has not been submitted to any other journal. This work did not require approval by a bioethical committee.

Data availability. Data are available from the corresponding author (dalvarez@agrosavia.co) upon reasonable request.

Author contribution statement (CRediT): **D. E. Álvarez-Sánchez:** validation, investigation, resources, data curation. **L. López-Rivera:** writing-original draft, writing-review & editing and project administration. **X. López-Estrella:** methodology, investigation, writing-original draft. **E. Manso-Ordoñez:** methodology, investigation, writing-original draft. **J. Rodríguez-Valenzuela:** validation, visualization and writing-review.

REFERENCIAS

- Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, 10(7), p.1046. <https://doi.org/10.3390/agronomy10071046>
- Abraham, S., Huynh, C. and Vu, H., 2020. Classification of Soils into Hydrologic Groups Using Machine Learning. *Data*, 5(1), pp. 2-11. <https://doi.org/10.3390/data5010002>
- Acuña, A.M. and Molina, P.E.C., 2018. Desempeño funcional del boro en las plantas. *UDO Agrícola*, 7(1), pp. 1-14. <https://www.bioline.org.br/pdf?cg07002>
- Álvarez Sánchez, D.E., Gómez, E. and Ordoñez, H., 2019. Tipología de fincas productoras de arveja (*Pisum sativum* L.) en la subregión Sur de Nariño, Colombia. *Ciencia y Tecnología Agropecuaria*, 20(3), pp. 659-677. https://doi.org/10.21930/rcta.vol20_num3_art:1593
- Álvarez Sánchez, D.E., Arévalo Rodríguez, A., Benavides Martínez, I.F., Salazar González, C. and Betancourth García, C., 2023. Use of Trained Convolutional Neural Networks for Analysis of Symptoms Caused by *Botrytis fabae* Sard. *Revista de Ciencias Agrícolas*, 40(1), pp. e1198. <https://doi.org/10.22267/rcia.20234001.198>
- Alzate, J.G., Benavides-Cardona, C.A. and Sánchez, D.E.Á., 2022. Aplicación de Urea, DAP y KCl sobre almacigos de café (*Coffea arabica* L.) de la variedad Castillo. *Revista de Investigación Agraria y Ambiental*, 13(1), pp. 35-55. <https://doi.org/10.22490/21456453.4297>
- Apráez, E., Gálvez, A. and Apráez, J., 2019. Factores edafoclimáticos en la producción y calidad del pasto Saboya (*Holcus lanatus* L.) en el Altiplano de Nariño. *Revista de Ciencias Agrícolas*, 36(1), pp. 16-32. <https://doi.org/10.22267/rcia.193601.95>
- Aworka, R., Cedric, L.S., Adoni, W.Y.H., Zoueu, J.T., Mutombo, F.K., Kimpolo, C.L.M., Nahhal, T. and Krichen, M., 2022. Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology*, 2, pp. 100-148. <https://doi.org/10.1016/j.jatech.2022.100048>
- Barbier, M., 2019. Biodisponibilidad de boro derivado de tres fuentes de boro en dos tipos de suelo. *Palmas*, 40 (Especial), pp. 158-167. <https://publicaciones.fedepalma.org/index.php/palmas/article/view/13045>
- Bayat, H., Asghari, S., Rastgou, M. and Sheykhzadeh, G.R., 2020. Estimating Proctor parameters in agricultural soils in the Ardabil plain of Iran using support vector machines, artificial neural networks and regression methods. *Catena*, 189, pp. 10-44. <https://doi.org/10.1016/j.catena.2020.104467>
- Becerra-Rondón, A., Ducati, J. and Haag, R., 2022. Satellite-based estimation of NO₂ concentrations using a machine-learning

- model: A case study on Rio Grande do Sul, Brazil. *Atmosfera*, 37, pp. 175- 190. <https://doi.org/10.20937/ATM.53116>
- Bouasria, A., Namr, K.I., Rahimi, A. and Ettachfani, E.M., 2020. Estimate soil organic matter from remote Sensing data by using statistical predictive models. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 1106-1115. https://doi.org/10.1007/978-3-030-90633-7_98
- Camacho, Y.M., Acosta, D.F., López, C.A., Cely, G.E. and Serrano, P.A., 2020. Transformación electroquímica de un Andisol en relación con la dinámica del fósforo. *Bioteología en el Sector Agropecuario y Agroindustrial*, 18(2), pp. 94-102. [https://doi.org/10.18684/BSAA\(18\)94-102](https://doi.org/10.18684/BSAA(18)94-102)
- Castro-Rincón, E., Cardona-Iglesias, J.L., Hernández-Oviedo, F., Valenzuela-Chiran, M. and Avellaneda-Avellaneda, Y., 2019. Evaluación de tres cultivares de *Lolium perenne* L. con vacas lecheras, en el trópico alto de Nariño-Colombia. *Pastos y Forrajes*, 42(2), pp. 161-170. <https://www.redalyc.org/journal/2691/269161217009/html/>
- Chen, S., Webb, G.I., Liu, L. and Ma, X., 2020. A novel selective naïve Bayes algorithm. *Knowledge- Based Systems*, 192, pp. 10-53. <https://doi.org/10.1016/j.knosys.2019.105361>
- Choi, E.Y., Park, H.I., Ju, J.H. and Yoon, Y.H., 2015. Boron availability alters its distribution in plant parts of tomato. *Horticulture, Environment, and Biotechnology*, 56(2), pp. 145-151. <https://doi.org/10.1007/s13580-015-0044-y>
- Culchac, L.Y., Estrada, J.S. and Ordóñez, H.R., 2021. Cuantificación de bacterias nitrificantes en un suelo Typic melanudands en tres condiciones de uso de suelo en Pasto, Nariño, Colombia. *Ciencia y Tecnología Agropecuaria*, 22(2), pp. 1-15. https://doi.org/10.21930/rcta.vol22_num2_art:1424
- El-Sayed, M.A., Abd-Elazem, A.H., Moursy, A.R.A., Mohamed, E.S., Kucher, D.E. and Fadl, M.E., 2023. Integration Vis-NIR spectroscopy and artificial intelligence to predict some soil parameters in arid region: A case study of Wadi Elkobaneyya, South Egypt. *Agronomy*, 13(3), pp. 35-49. <https://doi.org/10.3390/agronomy13030935>
- Ennaji, O., Vergütz, L. and El Allali, A., 2023. Machine learning in nutrient management: A review. *Artificial Intelligence in Agriculture*, 9, pp. 1-11. <https://doi.org/10.1016/j.aiia.2023.06.001>
- Galupino, J. and Dungca, J., 2022. Machine learning models to generate a subsurface soil profile: A case of Makati City, Philippines. *GEOMATE Journal*, 23(95), pp. 57-64. <https://geomatejournal.com/geomate/article/view/3029>
- Instituto Geográfico Agustín Codazzi, 2004. Estudio general de suelos y zonificación de tierras departamento de Nariño. IGAC. 2017. Guía de muestreo. <https://antiguo.igac.gov.co>
- Ismaili, M., Krimissa, S., Namous, M., Htitiou, A., Abdelrahman, K., Fnais, M. S., Lhissou, R., Eloudi, H., Faouzi, E. and Benabdelouahab, T., 2023. Assessment of soil suitability using Machine Learning in arid and semi-arid regions. *Agronomy*, 13(1), pp. 165-178. <https://doi.org/10.3390/agronomy13010165>
- Kaushik, K., Bhardwaj, A., Dahiya, S., Maashi, M. S., Al Moteri, M., Aljebreen, M. and Bharany, S., 2022. Multinomial Naive Bayesian classifier framework for systematic analysis of smart IoT devices. *Sensors*, 22(19), pp. 1-16. <https://doi.org/10.3390/s22197318>
- Kebonye, N.M., Agyeman, P.C. and Biney, J.K.M., 2023. Optimized modelling of country wide soil organic carbon levels via an interpretable decision tree. *Smart Agricultural Technology*, 3, pp. 100-106. <https://doi.org/10.1016/j.atech.2022.100106>
- Lagos, J. and Huertas, J., 2019. Efectos de una enmienda orgánica sobre las propiedades fisicoquímicas de un Andisol, en Túquerres, Nariño, Colombia. *Agro sur*, 47(3), pp. 39-54. https://doi.org/10.4206/agrosur_2019.v47n3-05
- Malik, P., Sengupta, S. and Jadon, J.S., 2021. Comparative analysis of soil properties to predict fertility and crop yield using Machine Learning Algorithms. 11th International Conference on Cloud Computing, *Data Science & Engineering (Confluence)*, pp. 1004-1007. <https://doi.org/10.1109/Confluence51648.2021.9377147>

2

- Mallah, S., Delsouz Khaki, B., Davatgar, N., Scholten, T., Amirian-Chakan, A., Emadi, M., Kerry, R., Mosavi, A.H. and Taghizadeh-Mehrjardi, R., 2022. Predicting soil textural classes using Random Forest models: Learning from imbalanced dataset. *Agronomy*, 12(11), pp. 1-16.
<https://doi.org/10.3390/agronomy12112613>
- Marcillo-Paguay, C.A., Benavides-Cardona, C.A., Ramos-Zambrano, H.S. and Romero, J.V., 2022. Respuesta agronómica y económica de papa (*Solanum tuberosum* subsp. Andigena) a la fertilización diferencial en cuatro ambientes de Nariño, Colombia. *Revista Colombiana de Ciencias Hortícolas*, 16(2), pp. 1-11.
<https://doi.org/10.17584/rcch.2022v16i2.13559>
- Mason-Jones, K., Schmücker, N. and Kuzyakov, Y., 2018. Contrasting effects of organic and mineral nitrogen challenge the N-Mining Hypothesis for soil organic matter priming. *Soil Biology and Biochemistry*, 124, pp. 38-46.
<https://doi.org/10.1016/j.soilbio.2018.05.024>
- Medina, M., Arismendi, J. and Cerón-Muñoz, M., 2019. Descripción de las relaciones entre parámetros químicos en suelos de fincas lecheras del norte de Antioquia, Colombia. *Livestock Research for Rural Development*, 31(7), pp.#102.
<http://www.lrrd.org/lrrd31/7/ceron31102.htm>
- Meshram, V., Patil, K., Hanchate, D. and Ramkteke, S.D., 2021. Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, 1, pp. 1-11.
<https://doi.org/10.1016/j.aailsci.2021.100010>
- Mohd, M., Hashim, N., Abd, S. and Lasekan, O., 2022. Characterisation of pineapple cultivars under different storage conditions using infrared thermal imaging coupled with Machine Learning Algorithms. *Agriculture*, 12(7), pp.1013.
<https://doi.org/10.3390/agriculture12071013>
- Novoa, M. A., Miranda, D., & Melgarejo, L., 2018. Efecto de las deficiencias y excesos de fósforo, potasio y boro en la fisiología y el crecimiento de plantas de aguacate (*Persea americana*, cv. Hass). *Revista Colombiana de Ciencias Hortícolas*, 12(2), pp. 293-307.
<http://repositorio.uptc.edu.co/handle/001/293>
- Pant, H., Lohani, M.C. and Bhatt, A., 2019. Impact of physico-chemical properties for soils type classification of OAK using different machine learning techniques. *International Journal of Computer Applications*, 177(17), pp. 38-44.
<http://doi.org/10.5120/ijca2019919617>
- Rajković, D., Marjanović Jeromela, A., Pezo, L., Lončar, B., Zanetti, F., Monti, A., & Kondić Špika, A., 2021. Yield and quality prediction of winter rapeseed—Artificial neural network and random forest models. *Agronomy*, 12(1), p.58.
<https://doi.org/10.3390/agronomy12010058>
- Ramírez, C.A., 2020. Aplicación del Machine Learning en agricultura de precisión. *Revista CINTEX*, 25(2), pp. 14-27.
<https://doi.org/10.33131/24222208.356>
- Rodríguez-Rodríguez, D.M., Checa-Coral, Ó.E., Ruiz-Eraso, H., Muriel-Figueroa, J. and Yepes-Chamorro, B., 2022. Niveles de fertilización con calcio, magnesio y azufre en genotipos de arveja voluble (*Pisum sativum* L.) en Nariño. *Actualidad & Divulgación Científica*, 25(2), pp. 1-8.
<https://doi.org/10.31910/rudca.v25.n2.2022.1847>
- Soil Survey Staff., 1999. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. 2nd edition. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 436.
<https://doi.org/10.1111/j.1475-2743.2001.tb00008.x>
- Volverás-Mambuscay, B., Merchancano-Rosero, J.D., Campo-Quesada, J.M. and López-Rendón, J.F., 2020. Propiedades físicas del suelo en el sistema de siembra en wachado en Nariño, Colombia. *Agronomía Mesoamericana*, 31(3), pp. 743-760.
<https://doi.org/10.15517/am.v31i3.39233>
- Yepes, J., Oré, G., Alcántara, M.S., Hernandez-Figueroa, H.E. and Teruel, B., 2022. Classification of sugarcane yields according to soil fertility properties using supervised machine learning methods. *Engenharia Agrícola*, 42(5), pp. 1-9.
<https://doi.org/10.1590/1809-4430-eng.agric.v42n5e20210239/2022>
- Zhang, J., He, Y., Yuan, L., Liu, P., Zhou, X., &

- Huang, Y., 2019. Machine learning-based spectral library for crop classification and status monitoring. *Agronomy*, 9(9), p.496. <https://doi.org/10.3390/agronomy9090496>
- Zhu, J., Jin, X., Li, S., Han, Y. and Zheng, W., 2022. Prediction of soil available Boron content in Visible-Near-Infrared hyperspectral based on different preprocessing transformations and characteristic wavelengths modeling. *Computational Intelligence and Neuroscience*, 20(22), pp.1-16. <https://doi.org/10.1155/2022/9748257>
- Zolfaghari, M., Moradi, M., Moradi, G. and Taghizadeh-Mehrjardi, R., 2023. Machine Learning models for prediction of soil properties in the Riparian Forests. *Land*, 12(1), pp. 32. <https://doi.org/10.3390/land12010032>