Predicción de la capacidad de intercambio catiónico (CIC) en cultivos de aguacate empleando modelos Machine Learning

Prediction Cation Exchange Capacity (CEC) in avocado crops using Machine Learning

María Haidy Castaño Robayo, mhcastanor@libertadores.edu.co, Fundación Universitaria Los Libertadores

José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co, Fundación Universitaria Los Libertadores

RESUMEN

La gran demanda alimentaria ha conllevado a la implementación de prácticas agrícolas que permitan optimizar las propiedades del suelo y, por tanto, aumentar la producción de los cultivos. La capacidad de intercambio catiónico (CIC) es un indicador de la capacidad del suelo para retener e intercambiar nutrientes a las plantas y se relaciona con la fertilización de los cultivos. Un valor alto de CIC brinda mayor capacidad para disponer de los nutrientes mientras que un valor bajo indica una menor disponibilidad de intercambio de estos. Por lo tanto, al relacionarse la CIC con la capacidad de absorción de los nutrientes en el suelo se pueden proponer estrategias que reduzcan el gasto innecesario en insumos de fertilizantes y también evitar impactos ambientales. En este trabajo se propone emplear modelos supervisados de Machine Learning con el fin de predecir y establecer las variables que

influyen sobre la CIC en el cultivo de aguacate. Para esto, se emplea una base de datos proporcionada por la Corporación Colombiana de Investigación Agropecuaria - Corpoica y publicada en datos abiertos Colombia. Con los datos de departamento, municipio, pH y las variables transformadas de materia orgánica, conductividad eléctrica y acidez se entrenaron modelos de regresión con aprendizaje supervisado para pronosticar el valor del logaritmo en base 10 de CIC (CIC_log10) y se emplearon indicadores como el error porcentual absoluto medio (MAPE) para evaluar su desempeño. El Random Forest Regressor presentó las mejores métricas y permitió establecer que las variables CE_log10 y acidez_log10 tienen un impacto mayor sobre el pronóstico de CIC_log10. Los resultados sugieren que el modelo presenta una exactitud del 79% y un F1score del 67% cuando se predice la CIC y posiblemente se pueden realizar recomendaciones aproximadas sobre las estrategias de fertilización en cultivos de aguacate.

Palabras clave: cultivo, fertilización, capacidad de intercambio catiónico (CIC), aprendizaje supervisado

ABSTRACT

High food demand requires optimizing soil properties for better production. Cation exchange capacity (CEC) indicates the capacity of the soil to retain and exchange nutrients to plants and affects crop fertilization. A high CEC provides a greater capacity to dispose of nutrients while a low value indicates lower availability of nutrients. Therefore, the CEC helps to propose fertilization strategies to reduce costs and environmental effects. In this work, supervised Machine Learning models were proposed to establish the variables that influence

the CEC in the avocado crop. The study used a database provided by la Corporación Colombiana de Investigación Agropecuaria Corpoica and published in open data Colombia. With data from the department, municipality, pH and the transformed variables of organic matter, electrical conductivity and acidity, regression models were trained with supervised learning to predict the logarithm in base 10 of CEC. The mean absolute percentage error (MAPE) was used to evaluate their performance. Random Forest Regressor presented the best metrics and established that the variables CE_log10 and acidity_log10 have the greatest impact on the CIC_log10 forecast. When CEC is predicted the model showed an accuracy of 79% and F1 score of 67% and a rough recommendation on fertilization strategies in avocado crops can be made.

Keywords: crop, fertilization, cation exchange capacity (CEC), supervised learning,

INTRODUCCIÓN

El suelo constituye la base en la producción alimentaria al suministrar los nutrientes esenciales, agua, oxígeno y soporte a las raíces de las plantas que producen nuestro alimento. Además de producir alimentos nutritivos y de buena calidad, conservar los suelos sanos ayuda a mitigar el cambio climático, a mantener la biosfera por permitir la presencia de distintas especies de plantas y animales y también a almacenar agua (FAO, 2015). Considerando los beneficios de preservar suelos saludables, es importante realizar un manejo adecuado para su conservación, siendo una estrategia la aplicación adecuada de fertilizantes

al suelo que le permita obtener una cantidad óptima de nutrientes disponibles para la planta, al estimar características del suelo como su capacidad de intercambio.

La capacidad de intercambio catiónico (CIC) es una medida de la capacidad del suelo para absorber cationes y determina la cantidad de sitios disponibles para almacenar cationes en el suelo, los cuales pueden ser intercambiados por otros, convirtiéndose en cationes intercambiables necesarios en los procesos de nutrición de la planta. Los cationes más importantes en los procesos de intercambio catiónico son $Ca^{2+,}$ Mg^{2+}, K^{+} y Na^{+} debido a que se encuentran en mayores cantidades y constituyen las bases del suelo.

La CIC puede expresarse en unidades de cmol/kg que corresponde a centimoles por kilogramo de suelo o en meq/100g que hace referencia a miliequivalentes por 100g de suelo, ambas unidades son numéricamente iguales (Jaramillo Jaramillo, 2002).

La CIC es un indicador de la calidad del suelo y es de gran importancia para conocer el potencial de un suelo para reservar e intercambiar nutrientes, lo que afecta directamente la frecuencia y cantidad de fertilizantes aplicados(Khaledian et al., 2017). Por ejemplo, existen suelos agotados por deficiencia de nutrientes pero que tienen buena capacidad de almacenamiento requiriendo una mayor adición de abonos mientras que existen otros suelos sin nutrientes y además con baja capacidad de almacenarlos que requieren baja cantidad de abono, por lo que, la estrategia de fertilización difiere en estos suelos en relación con su CIC (INTAGRI, n.d.; Rengifo Mejía et al., 2020).

Algunas propiedades químicas como el pH, el contenido de materia orgánica, la conductividad eléctrica y la acidez intercambiable, se relacionan estrechamente con la fertilidad del suelo al influir sobre la asimilación, disponibilidad y retención de los nutrientes (Estrada-Herrera et al., 2017).

El pH que es una medida de la acidez o basicidad, indica la concentración de iones H^+ presentes en la disolución del suelo e influye en la disponibilidad de los nutrientes para las plantas y otros contaminantes inorgánicos. Generalmente, el valor del pH del suelo se encuentra entre 3,5 y 9,5 siendo los valores < 5,5 clasificados como muy ácidos y tienden a presentar intoxicación por aluminio o manganeso y suelos muy alcalinos > 8,5 tienden a presentar precipitación de hidróxidos de los cationes por la presencia del grupo OH- (FAO, 2020).

Por otro lado, la materia orgánica hace referencia a la fracción orgánica que posee el suelo y aporta nutrientes como nitrógeno, fósforo y azufre, pero también puede disminuir la disponibilidad de algunos otros por la formación de complejos con Cu, Mn, Zn entre otros. De manera general, la materia orgánica contribuye al incremento del valor del CIC (Jaramillo Jaramillo, 2020; Julca-Otiniano et al., 2006).

La conductividad eléctrica hace referencia a la cantidad de sales disponibles que son capaces de conducir corriente, depende del contenido de agua y de la presencia de iones intercambiables, además de ser una medida indirecta de la salinidad del suelo que es un fenómeno indeseable porque inhibe el crecimiento de las plantas. Debido a que los fertilizantes están compuestos por sales inorgánicas, la elección adecuada de fertilizante y su cantidad tendrán un efecto importante sobre la conductividad y la CIC(Redagrícola, 2017).

La acidez está relacionada con la presencia de aluminio e hidrógeno y se presenta a pH inferiores a 5,5. Esta propiedad se relaciona con la disponibilidad de los nutrientes, ya que, en suelos muy ácidos las posiciones de intercambio están ocupadas por iones Al^{3+} y H^+ disminuyendo el valor de CIC, además la presencia de Al^{3+} resulta ser tóxico para muchas plantas. Sin embargo, este efecto depende del tipo de cultivo ya que no todas las plantas toleran las mismas cantidades de aluminio en el suelo (Rengifo Mejía et al., 2020).

El cultivo de aguacate ha ganado gran importancia en Colombia en los últimos años debido a que en el país se cuenta con las condiciones agroclimáticas necesarias para su desarrollo, además, de ayudar a fortalecer la oferta exportadora de productos no tradicionales en el país y contribuyendo a la generación de empleos, ya que, se estima que cerca de 65000 personas se ven involucradas en la cadena productiva del aguacate. Colombia es el cuarto productor de aguacate en el ranking mundial con una producción de 550000 toneladas, la variedad Hass corresponde al 34% del total de área sembrada con aguacate del país y el 65% de su producción se queda en el mercado nacional y el 35% restante se entrega a mercados de exportación. Antioquia registra como el departamento con mayor producción a nivel nacional, seguido de Caldas y Tolima con 23%, 16% y 14% respectivamente (Minagricultura, 2021).

Considerando la importancia de la producción de aguacate actualmente en el país, es de gran interés el estudio de las propiedades químicas del suelo como la CIC, porque brinda información sobre su salud y permite proponer estrategias para su correcta fertilización, esto al considerar la normativa relacionada con los cultivos de aguacate que destaca la importancia de diseñar un plan de fertilización para la nutrición del cultivo basado en el análisis de suelo y los requerimientos de la especie sembrada (Resolución ICA 30021, 2017), ya que, una sobredosificación o subdosificación puede acarrear costos económicos y ambientales, como es reportado en literatura (Ruíz Arredondo, D., 2019).

Es así que, los métodos Machine learning representan una alternativa novedosa para determinar propiedades de los suelos (Syah et al., 2021), lo que permite la toma de decisiones con menores impactos sobre los ecosistemas y probablemente a una reducción de costos por

el uso adecuado de fertilizantes. Por esta razón, existen diversos trabajos que estiman propiedades del suelo a través de modelos de regresión, redes neuronales, entre otros métodos (Ghorbani et al., 2015; Seybold et al., 2005; Shabani & Norouzi Masir, 2015), encontrando resultados razonables en sus pronósticos, siendo estos modelos una herramienta valiosa que proporciona parámetros del suelo con datos fácilmente disponibles.

Considerando lo anteriormente expuesto, se plantea la formulación de modelos Machine Learning para predecir el valor de CIC en cultivos de aguacate a nivel nacional empleando propiedades químicas del suelo fáciles de medir como pH, conductividad eléctrica, contenido de materia orgánica y acidez intercambiable, esto con el fin de establecer relaciones entre estos parámetros y la CIC y probablemente proponer recomendaciones para una adecuada fertilización.

METODOLOGÍA

El procedimiento desarrollado comprende cuatro etapas principales; recopilación, preprocesamiento, análisis descriptivo y modelación de los datos. Los tratamientos y análisis se realizaron en Python empleando las librerías SweetViz, pycaret y sklearn.

- *Recopilación de datos*

La base de datos empleada en el presente trabajo comprende los resultados del análisis de laboratorio de suelos en Colombia. La información es proporcionada por la Corporación Colombiana de Investigación Agropecuaria - Corpoica con fecha de actualización del 11 de agosto del 2020, la cual, se encuentra disponible en datos abiertos Colombia a través del siguiente enlace:

https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Resultados-de-An-lisis-de-Laboratorio-Suelos-en-Co/ch4u-f3i5.

La versión original contiene 46700 registros con 33 variables que incluyen información sobre el tipo de cultivo, lugar de recolección y parámetros fisicoquímicos del suelo.

- *Preprocesamiento*

Para la depuración de los datos originales se seleccionó el cultivo de aguacate como el filtro principal. Se eliminaron las variables *numfila, estado, tiempo de establecimiento, topografía, drenaje, riego, fertilizantes aplicados, fecha de análisis, secuencial y las variables de los macro y micronutrientes*, resultando dos variables categóricas (*departamento y municipio*) y las demás numéricas (*capacidad de intercambio catiónico, materia orgánica, ph agua suelo, conductividad eléctrica y acidez*). Para los registros con valor de cero en *acidez* se realizó una imputación con la media. De esta manera, la base final abarca cinco categorías numéricas asociados a los parámetros fisicoquímicos del suelo con 1838 registros en total.

Finalmente, y al considerar el propósito de establecer relaciones entre las variables del suelo e intentar planear estrategias de fertilización, se estableció como variable objetivo la capacidad de intercambio catiónico (CIC) y como variables predictoras las medidas de pH, el porcentaje de materia orgánica , la conductividad eléctrica y la acidez intercambiable junto con el departamento y municipio. Sin embargo, se realizó la transformación de algunas variables numéricas con el logaritmo en base 10 para obtener datos más comparables entre sí, mejorar la suposición de normalidad y adecuar los datos para aplicar los modelos. En la **Tabla 1** se describe cada una de las variables con su unidad junto con las transformaciones empleadas en los modelos desarrollados en el trabajo.

- *Análisis descriptivo exploratorio*

A partir de esta etapa se emplearon las nuevas variables para realizar la estimación de CIC_log10. El conjunto de datos se analizó empleando histogramas de las variables y su relación con la CIC_log10 , mapas de correlación y parámetros como media, mediana y cuartiles.

Tabla 1. Identificación de las variables y transformaciones

Variable	Descripción	Transformación	Nueva variable
Departamento	Ubicación del suelo por departamento, Antioquía, Cundinamarca, Valle del cauca, Cauca entre otros.	Ninguna	departamento
Municipio	Ubicación del suelo por municipio, Rionegro, Bogotá, entre otros.	Ninguna	municipio
Capacidad de intercambio catiónico (CIC)	**Variable objetivo.** Indica la disponibilidad de nutrientes. Se expresa en cmol por Kg de suelo	Se aplica el log 10	CIC_log10
materia orgánica	Contenido de carbono orgánico expresado en %	Se aplica el log 10	MO_log10
ph agua del suelo	Grado de acidez o alcalinidad del agua del suelo (sin unidades)	Ninguna	ph_suelo
Conductividad eléctrica	Relación de iones presentes provenientes de las sales en ds/m	Se aplica el log 10	CE_log10
Acidez	Contenido de iones Al^{+3} y H^{+} en la solución del suelo expresado en cmol por Kg de suelo	Se aplica el log 10	acidez_log10

- *Modelación*

Se realizó la separación de los datos en entrenamiento (80%; n=1470) y prueba (20%; n= 368). El desempeño de los modelos se evaluó usando las métricas de RMSE, MAE, R^2 y MAPE, siendo esta última la empleada para elegir el mejor modelo de acuerdo con la siguiente fórmula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| x\ 100 \qquad \textbf{Ec. 1}$$

Se implementó como modelo base una regresión lineal entre la variable predictora acidez_log10 y la variable objetivo CIC_log10. Adicionalmente, se desarrollaron modelos Machine Learning y se seleccionó el modelo con más bajo MAPE, además, se examinaron los residuales, la predicción del error y la importancia de las variables predictoras. Posteriormente, se realizaron pronósticos con los datos de prueba para estimar la variable sin transformar CIC y se construyó una matriz de confusión para evaluar su desempeño.

RESULTADOS

La **Figura 1** muestra la relación entre las variables del suelo evaluadas en el trabajo. De manera general, se observa una importante correlación negativa entre el ph_suelo y acidez_log10 mientras que las correlaciones positivas importantes se presentan entre MO_log10 y CE_log10, para el caso de CIC_log10 se evidencia mayor relación con CE_log10 y acidez_log10.

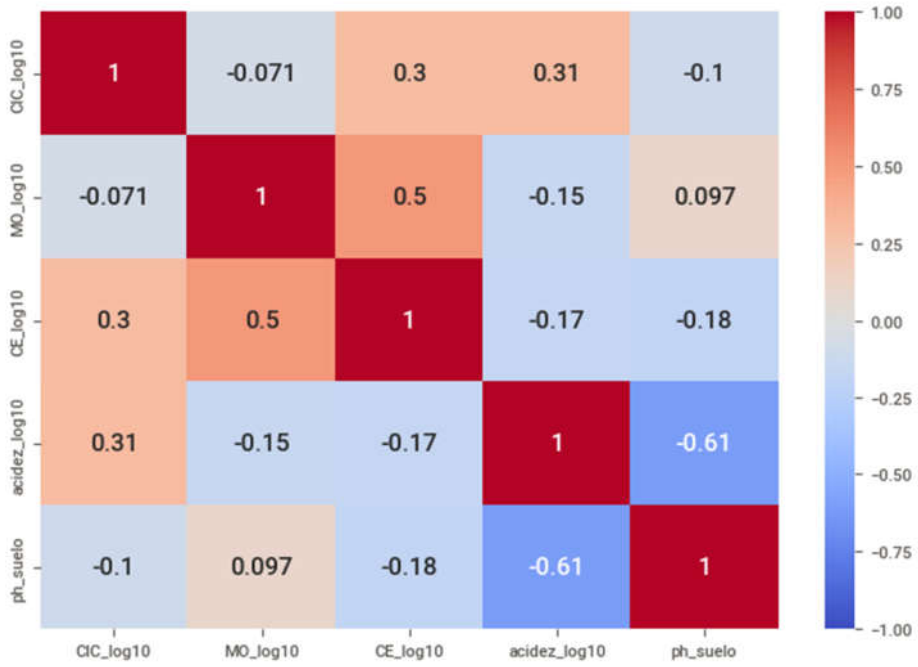

Figura 1. Matriz de correlación entre las variables transformadas del suelo

En la **Figura 2** ae aprecia que los valores de CIC_log10se encuentran en el rango de -0,01 y 1,4, con un valor medio de 0,63 y siendo el valor de 0,61 el que presenta la mayor frecuencia. De acuerdo a los resultados, los suelos estudiados evidencian un intervalo amplio desde valores bajos de CIC_log10 <0,70 (CIC< 5 cmol/Kg suelo) hasta valores altos de CIC_log10>1,30 (CIC >20 cmol/Kg suelo) según lo reportado (Rengifo Mejía et al., 2020).

El análisis descriptivo exploratorio permitió establecer las tendencias de las variables predictoras con la variable objetivo (CIC_log10), además de conocer los rangos principales en los que se encuentran estas propiedades.

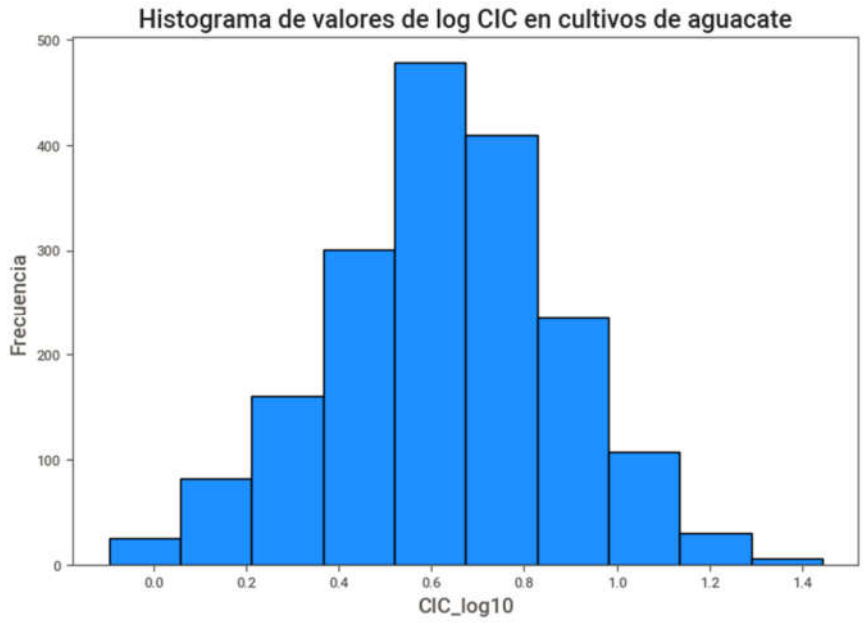

Figura 2. Distribución de los valores de CIC_log10

En la **Figura 3** se observa que el ph_suelo se encuentra en el rango de 3,80 a 5,52, el cual difiere del rango óptimo del pH reportado para el cultivo de aguacate entre 5,5 y 7, 5, (Bisonó SM, 2008), además se aprecia una tendencia inversa con CIC_log10, en donde esta última presenta valores grandes cuando se tienen ph_suelo bajos (< 4,25) y viceversa. Sin embargo, el ph_suelo no presenta una correlación tan importante con CIC_log, de acuerdo con los resultados del mapa de calor mostrados en la Figura 1.

Por otro lado, se observan valores muy diversos para MO_log10 que se encuentran en el rango de -0,70 y 1,64, pero su correlación con CIC_log10 no es tan significativa (-0.071).

Los rangos de CE_log10 y acidez_log10 corresponden a -1,70 a 0,60 y -1.40 a 1,16 respectivamente. La Figura 3 hace más evidente la correlación positiva de estas variables con la CIC_log10, ya que, la CIC_log10 exhibe los mayores valores cuando se tienen suelos con valores de CE_log10 de 0,60 y acidez_log10 de 1,16.

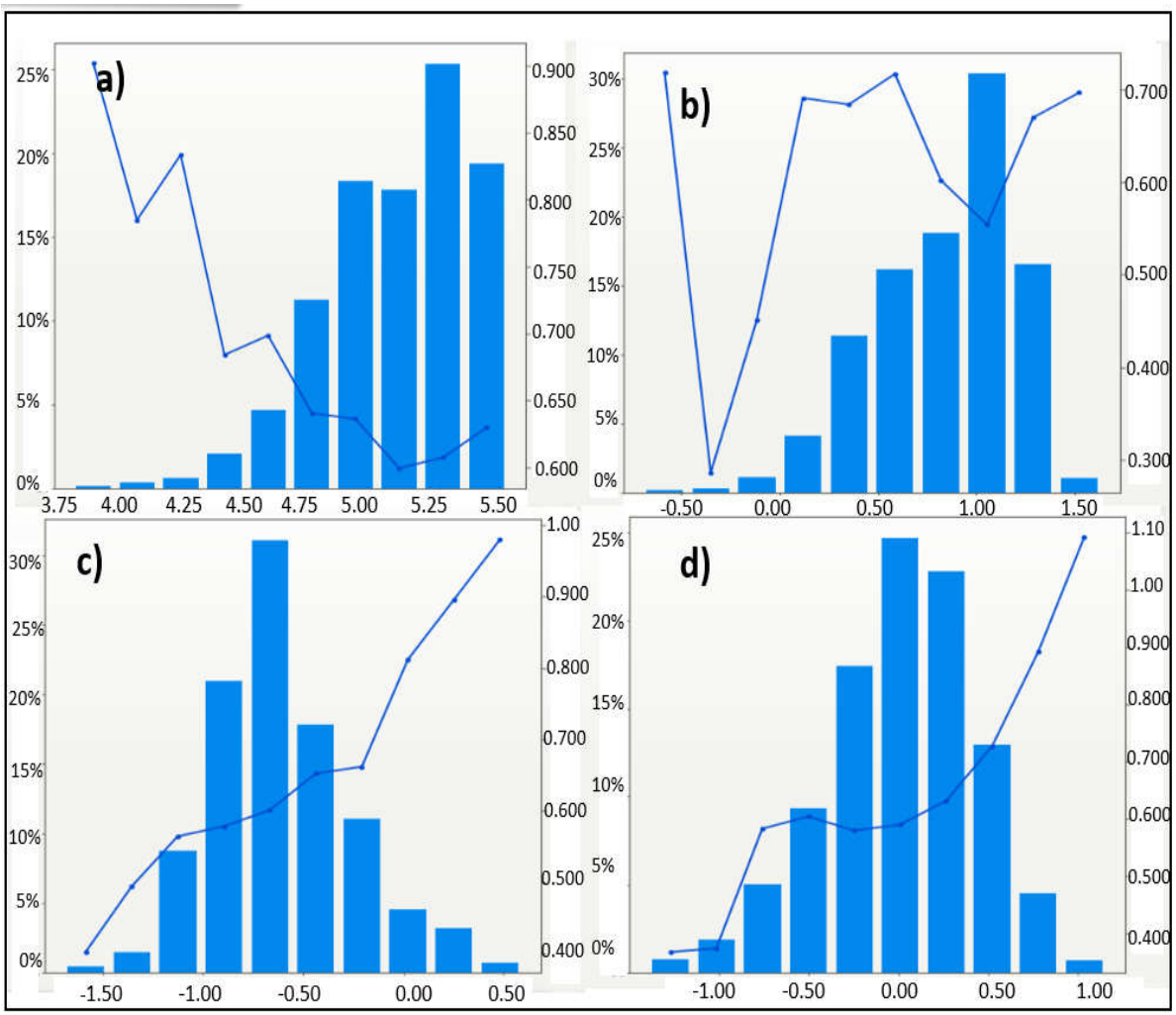

Figura 3. Relación entre la variable objetivo CIC_log10 (—●—) y las variables predictoras **a)** ph_suelo, **b)** MO_log10, **c)** CE_log10, **d)** acidez_log10

De acuerdo con los resultados encontrados en el análisis exploratorio, se determinó que la acidez_log10 exhibe la mayor correlación con CIC_log10 (0,31),por esta razón, se planteó un modelo base de regresión lineal únicamente entre estas dos variables para comparar su rendimiento con modelos Machine Learning que si involucran todas las entradas predictoras categóricas y numéricas (departamento, municipio, ph_suelo, MO_log10, CE_log10 y

acidez_log10). La **Tabla 2** muestra las métricas de desempeño del modelo base y de los cinco mejores modelos Machine Learning encontrados en el desarrollo del presente trabajo.

Tabla 2. Comparación del desempeño de los modelos

Modelo	MAE	RMSE	R^2	MAPE
Regresión lineal (modelo base)	**0,179**	**0,234**	**0,108**	**1,08**
Extra Trees Regressor	**0,110**	**0,151**	**0,620**	**0,417**
Random Forest Regressor	**0,113**	**0,155**	**0,596**	**0,421**
Gradient Boosting Regressor	0,121	0,160	0,568	0,471
Light Gradient Boosting Machine	0,122	0,167	0,513	0,459

El modelo base presenta un MAPE mayor en comparación a los demás modelos Machine Learning, indicando un bajo ajuste en el pronóstico de CIC_log10 por este método. Por su parte, Extra Trees Regressor y Random Forest Regressor exhiben los menores valores de MAE,RMSE y MAPE de todos los modelos evaluados (anexo A1), considerando solo el MAPE, se establece que el porcentaje de error entre las predicciones y los valores reales del CIC_log10 de estos dos modelos se encuentran alrededor del 42%.

Al evaluar los residuales y la predicción del error de los modelos Extra Trees y Random Forest Regressor, se evidenció un mayor sobreajuste con el Extra Trees Regressor, razón por la cual, se eligió el Random Forest Regressor como el algoritmo a emplear para realizar el pronóstico del CIC_log10 en los datos del cultivo de aguacate al considerar un mejor comportamiento en los residuales (anexo A2) . La **Figura 4** muestra el resultado del error usando este modelo.

De la figura se infiere que hay un sobreajuste a valores bajos, pero existe una subestimación a valores altos, por ejemplo, para un valor de CIC_log10 de 0,0 la predicción es cercana a 0,3 mientras que para 1,0 la predicción se encuentra por debajo de este valor; los resultados evidencian una deficiencia del modelo en generalizar con datos nuevos cuando se tienen bajos y altos valores de CIC_log10.

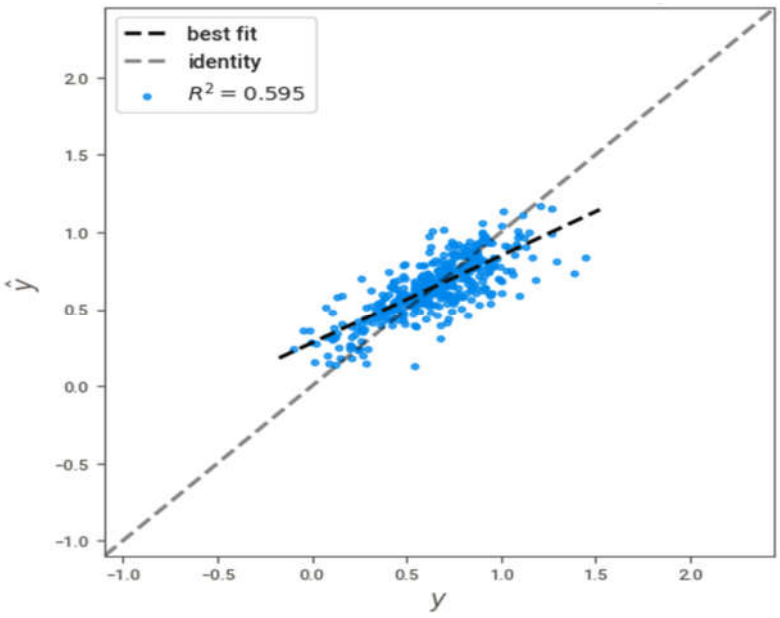

Figura 4. Predicción del error en Random Forest Regressor

La importancia de las categorías utilizadas para la predicción usando el modelo de aprendizaje el Random Forest Regressor aparece en la **Figura 5**. Los parámetros de entrada numéricos que más contribuyen son CE_log10 y acidez_log10 mientras que MO_log10 y ph_suelo parecen tener un menor impacto. En cuanto a las variables categóricas, de los 31 departamentos evaluados en este trabajo, Cundinamarca y Antioquia presentan el mayor efecto. La tendencia general es encontrar valores altos de CIC_log10 cuando se tienen valores altos de cada parámetro de entrada, sin embargo, en MO_log10 y

departamento_ANTIOQUIA, la tendencia es contraria, con valores bajos de estas variables la CIC_log10 es mayor.

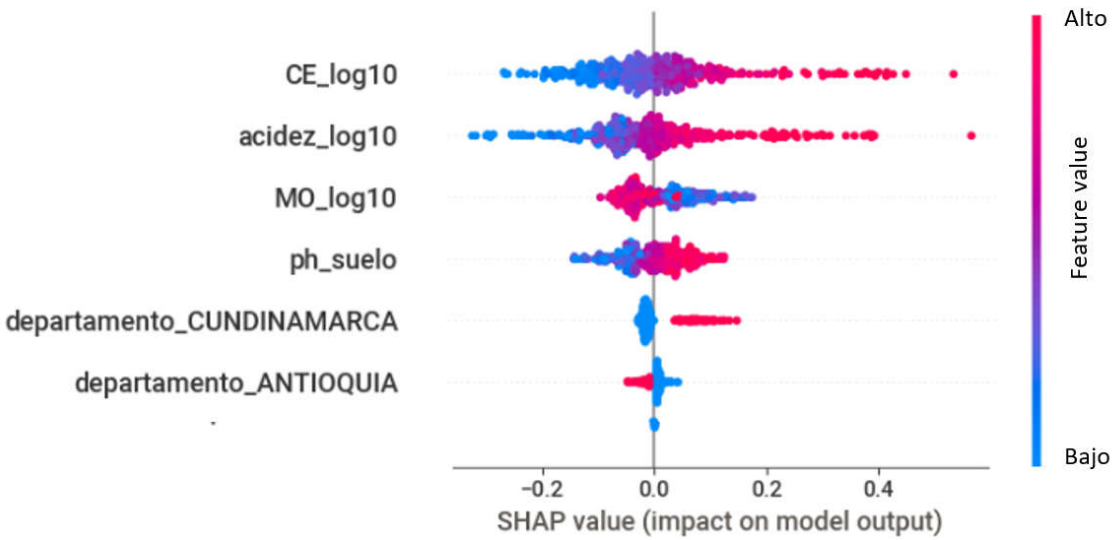

Figura 5. Interpretación del modelo empleando variables transformadas del análisis de suelo en cultivo de aguacate

Para analizar mejor el desempeño del modelo en la predicción de la CIC en cultivos de aguacate con datos diferentes a los de entrenamiento, se realizó una estimación utilizando los datos de prueba (n=368), con la condición de "aplicar más fertilizante" de 28,37kg/ha en contenido de Nitrógeno cuando se tiene una CIC media/alta (>5cmol/Kg suelo), de acuerdo con las recomendaciones encontradas en la localidad de Morales (Cauca) para la aplicación de fertilizantes en suelos con CIC bajas <5cmol/Kg (Corpoica, 2017). Los resultados de este ensayo empleando el modelo Random Forest Regressor son mostrados en la matriz de confusión de la **Figura 6** encontrando un valor de Accuracy del 79% , una precisión del 70%, un Recall del 65% y F1 score del 67% en el pronóstico.

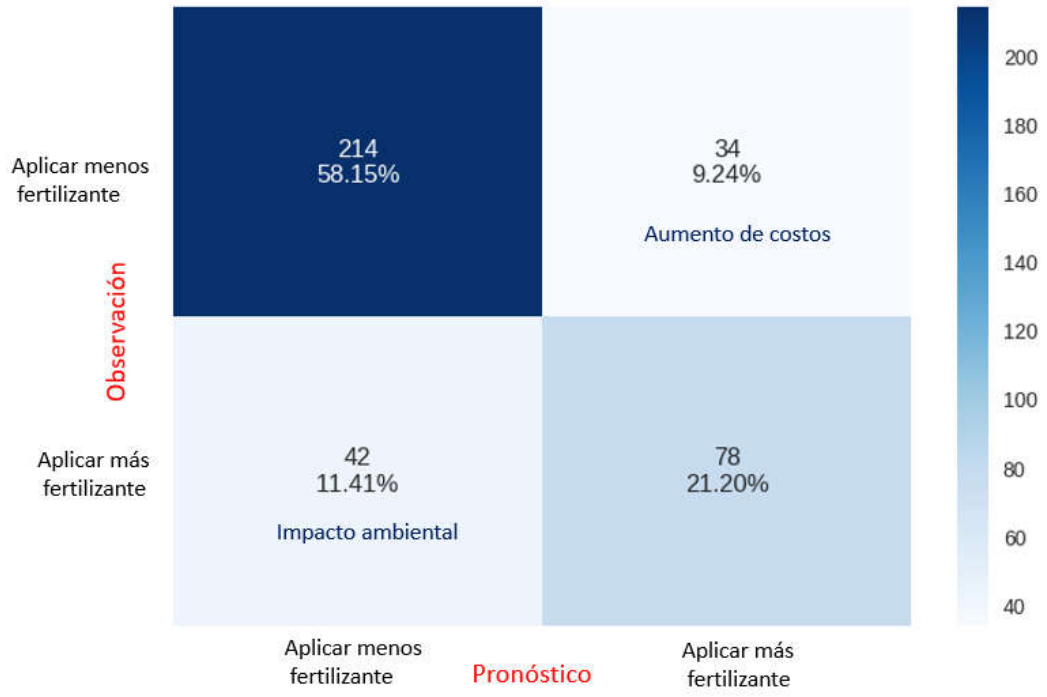

Figura 6. Matriz de confusión del modelo con los datos de prueba

DISCUSIÓN DE RESULTADOS

Uno de los propósitos del presente trabajo era establecer relaciones entre algunas propiedades del suelo con la CIC. Los resultados arrojaron que el ph_suelo y el MO_log10 presentan la menor correlación y menor efecto sobre el valor de CIC_log10. Se espera que a ph_suelo muy ácidos (< 5,5) ocurran efectos negativos sobre la actividad microbiana de la materia orgánica y el desarrollo de las plantas que pueden afectar la disponibilidad de nutrientes en suelo (Corpoica, 2017), sin embargo, si se considera el bajo impacto que tienen estas variables predictoras sobre la variable objetivo se disminuyen estos efectos adversos sobre el CIC del cultivo.

Con ayuda del Random Forest Regressor se identificó que de las variables estudiadas la CE_log10 es la más influyente en el pronóstico, lo que se relaciona con el efecto de la conductividad eléctrica sobre la calidad y fertilidad de los cultivos. Entonces se espera que a mayor valor de CE_log10 sea mayor la cantidad de cationes intercambiables, por lo que, se requerirá una mayor adición de fertilizantes al suelo. No obstante, para cultivos de aguacate se recomienda trabajar con valores de CE_log10 menores a 0,48 para evitar efectos tóxicos por cloruro de sodio como la absorción limitada de nutrientes y la presencia de antagonismos entre los iones (Rengifo Mejía et al., 2020).

Empleando los modelos de aprendizaje supervisado se encontraron valores más bajos de MAPE indicando una mejora en la exactitud de la predicción en comparación con el modelo base. Particularmente, el Random Forest Regressor es un algoritmo muy empleado en la determinación de parámetros de suelos (Garnaik et al., 2022; Makungwe et al., 2021), en relación con los resultados obtenidos este modelo presenta el menor error en las predicciones (MAPE=42%) al disminuir el sobreajuste, ya que, los bosques aleatorios se conforman por un conjunto de árboles decisión que hacen pronósticos a partir de la combinación de los resultados determinados en cada árbol de manera individual (Alvear, 2018).

En la evaluación del desempeño del Random Forest Regressor con los datos de prueba para pronosticar la adición de más o menos fertilizante (28,37kg/ha), se encontró que en el 58,15% de los casos el modelo recomienda aplicar menos fertilizante mientras que en el 21,20% recomienda aplicar más fertilizante mostrando que en un 79% el modelo realizó las predicciones correctamente. Sin embargo, el 9,24% de los casos corresponden a falsos positivos, es decir, el modelo pronóstica aplicar más fertilizante cuando realmente se debe aplicar menos, este pronóstico equivocado acarrearía un aumento en los costos, tal y como se evidenció en un cultivo de aguacate en el municipio de Abejorral -Antioquia, en donde la

aplicación de 30000g de fertilizante por árbol aumenta la inversión en $76984frente a los 460g de fertilizante que se requiere con un gasto de $618 (Ruíz Arredondo, D., 2019). Asimismo, la predicción de falsos negativos que corresponde al 11,41% constituye un riesgo ambiental porque el modelo pronóstica aplicar menor cantidad de fertilizante cuando realmente se debe aplicar más, ralentizando el crecimiento de las plantas, afectando directamente la calidad del suelo y la producción del fruto, tal y como se evidenció en cultivos de uchuva en donde la deficiencia de boro afectó el crecimiento de las ramas productivas, disminuyó el tamaño de los frutos y en consecuencia su producción (Fischer. et al., 2008).

A la luz de los resultados encontrados en el presente trabajo, se resalta que el modelo de aprendizaje supervisado propuesto presenta algunas limitaciones dado que se deben considerar variables estructurales del suelo, condiciones climáticas y además conocer el tipo de fertilizante aplicado para realizar una mejores estimaciones y proponer estrategias correctas de fertilización, sin embargo, estos resultados son una primera aproximación para determinar la CIC con variables predictoras fácilmente disponibles y pueden ser el punto de partida para proponer modelos Machine Learning incluyendo otras propiedades del suelo con mejores desempeños que contribuyan a mejorar las prácticas agrícolas.

CONCLUSIONES

Las variables CE_log10 y acidez _log10 presentan la mayor correlación con CIC_log10 y mayor impacto en el modelo mientras que el ph_suelo y MO_log10 presentan una menor relación.

El modelo Random Forest Regressor fue seleccionado para realizar los pronósticos como consecuencia de sus mejores métricas y disminución del sobreajuste en comparación con modelos de regresión lineal y Extra Trees Regressor.

Emplear la condición de "aplicar más o menos fertilizante" en suelos permitió analizar el desempeño en el pronóstico del Random Forest Regressor con datos reales y, establecer los posibles impactos ambientales y económicos que pueden presentarse en cultivos de aguacate si el modelo no predice correctamente, como lo es la acumulación de fertilizantes y los sobrecostos cuando se adiciona mayor cantidad de la necesaria.

REFERENCIAS

Alvear, J. O. (2018). *Ensambladores: Random Forest*. https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

Bisonó SM, H. J. (2008). *Guía tecnológica sobre el cultivo del aguacate*.

Corpoica. (2017). *Criterios para la definición de planes de fertilización en el cultivo de aguacate Hass con un enfoque tecnificado*.

Estrada-Herrera, I. R., Hidalgo-Moreno, C., Guzman-Plazola, R., Almaraz Suarez, J. J., Navarro-Garza, H., & Etchevers-Barra, J. D. (2017). Indicadores de calidad de suelo para evaluar su fertilidad. *Agrociencia*, *51*, 813–831.

FAO. (2015). *Los suelos sanos son la base para la producción de alimentos saludables*.

FAO. (2020). *Propiedades Químicas*. https://www.fao.org/soils-portal/soil-survey/clasificacion-de-suelos/sistemas-numericos/propiedades-quimicas/es/

Fischer., G., S., J., J., F., & M., F. E. (2008). Efecto de la deficiencia de N, P, K, Ca, Mg y B en componentes de producción y calidad de la uchuva (Physalis peruviana L.). *Agronomía Colombiana*, *26*, 389–398.

Garnaik, S., Samant, P. K., Mandal, M., Mohanty, T. R., Dwibedi, S. K., Patra, R. K., Mohapatra, K. K., Wanjari, R. H., Sethi, D., Sena, D. R., Sapkota, T. B., Nayak, J., Patra, S., Parihar, C. M., & Nayak, H. S. (2022). Untangling the effect of soil quality on rice productivity under a 16-years long-term fertilizer experiment using conditional

random forest. *Computers and Electronics in Agriculture*, *197*, 106965. https://doi.org/https://doi.org/10.1016/j.compag.2022.106965

Ghorbani, H., Kashi, H., Moghadas, N. H., & Emamgholizadeh, S. (2015). Estimation of Soil Cation Exchange Capacity using Multiple Regression, Artificial Neural Networks, and Adaptive Neuro-fuzzy Inference System Models in Golestan Province, Iran. *Communications in Soil Science and Plant Analysis*, *46*(6), 763–780. https://doi.org/10.1080/00103624.2015.1006367

Resolución ICA 30021, (2017).

INTAGRI. (n.d.). *La Capacidad de Intercambio Catiónico del Suelo*. https://www.intagri.com/articulos/suelos/la-capacidad-de-intercambio-cationico-del-suelo

Jaramillo Jaramillo, D. F. (2002). Los fenómenos de intercambio iónico. In *Introducción a la ciencia del suelo* (p. 321).

Jaramillo Jaramillo, D. F. (2020). La materia orgánica del suelo. In *ntroducción a la ciencia del suelo* (p. 417).

Julca-Otiniano, A., Meneses-Florián, L., Blas-Sevillano, R., & Bello-Amez, S. (2006). LA MATERIA ORGÁNICA, IMPORTANCIA Y EXPERIENCIA DE SU USO EN LA AGRICULTURA. *Idesia (Arica)*, *24*, 49–61.

Khaledian, Y., Brevik, E. C., Pereira, P., Cerdà, A., Fattah, M. A., & Tazikeh, H. (2017). Modeling soil cation exchange capacity in multiple countries. *CATENA*, *158*, 194–200. https://doi.org/https://doi.org/10.1016/j.catena.2017.07.002

Makungwe, M., Chabala, L. M., Chishala, B. H., & Lark, R. M. (2021). Performance of linear mixed models and random forests for spatial prediction of soil pH. *Geoderma*, *397*, 115079. https://doi.org/https://doi.org/10.1016/j.geoderma.2021.115079

Minagricultura. (2021). *Cadena productiva Aguacate*. Ministerio de Agricultura y Desarrollo Rural. https://sioc.minagricultura.gov.co/Aguacate/Pages/Documentos.aspx

Redagrícola. (2017). *Conductividad eléctrica y salinidad*. https://www.redagricola.com/cl/conductividad-electrica-salinidad/

Rengifo Mejía, P. A., Londoño Zuluaga, José David Diez Moreno, D., & Vásquez Yepes, G. E. (2020). Conceptos de fertilización para el cultivo de aguacate. *Servicio Nacional de Aprendizaje (SENA)*, 55.

Ruíz Arredondo, D., & R. M. (2019). Optimización de la fertilización del cultivo de aguacate CV.´HASS´ (Persea Americana Mill). *Encuentro Sennova Del Oriente Antioqueño*, *4*(1).

Seybold, C. A., Grossman, R. B., & Reinsch, T. G. (2005). Predicting Cation Exchange Capacity for Soil Survey Using Linear Models. *Soil Science Society of America Journal*, *69*(3), 856–863. https://doi.org/https://doi.org/10.2136/sssaj2004.0026

Shabani, A., & Norouzi Masir, M. (2015). Predicting Cation Exchange Capacity by Artificial Neural Network and Multiple Linear Regression Using Terrain and Soil Characteristics. *Indian Journal of Science and Technology*, *8*. https://doi.org/10.17485/ijst/2015/v8i28/83328

Syah, R., Al-Khowarizmi, A., Elveny, M., & Khan, A. (2021). Machine learning based simulation of water treatment using LDH/MOF nanocomposites. *Environmental Technology & Innovation*, *23*, 101805. https://doi.org/https://doi.org/10.1016/j.eti.2021.101805

ANEXOS

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.1160	2.400000e-02	0.1544	6.105000e-01	0.0955	0.3462	1.011
rf	Random Forest Regressor	0.1186	2.550000e-02	0.1591	5.877000e-01	0.0996	0.3712	1.006
lightgbm	Light Gradient Boosting Machine	0.1190	2.610000e-02	0.1615	5.718000e-01	0.1007	0.3573	0.108
gbr	Gradient Boosting Regressor	0.1242	2.680000e-02	0.1633	5.665000e-01	0.1031	0.3999	0.299
knn	K Neighbors Regressor	0.1357	3.160000e-02	0.1772	4.898000e-01	0.1098	0.4411	0.073
ridge	Ridge Regression	0.1391	3.300000e-02	0.1811	4.659000e-01	0.1117	0.3981	0.019
br	Bayesian Ridge	0.1400	3.310000e-02	0.1816	4.635000e-01	0.1122	0.4165	0.051
huber	Huber Regressor	0.1408	3.510000e-02	0.1867	4.304000e-01	0.1140	0.3498	0.170
omp	Orthogonal Matching Pursuit	0.1449	3.540000e-02	0.1877	4.263000e-01	0.1162	0.4004	0.018
ada	AdaBoost Regressor	0.1532	3.700000e-02	0.1919	4.032000e-01	0.1217	0.5690	0.208
par	Passive Aggressive Regressor	0.1593	4.000000e-02	0.1994	3.526000e-01	0.1250	0.5077	0.029
dt	Decision Tree Regressor	0.1606	5.000000e-02	0.2229	1.915000e-01	0.1399	0.3882	0.029
lasso	Lasso Regression	0.1958	6.270000e-02	0.2498	-8.400000e-03	0.1579	0.7616	0.016
en	Elastic Net	0.1958	6.270000e-02	0.2498	-8.400000e-03	0.1579	0.7616	0.017
llar	Lasso Least Angle Regression	0.1958	6.270000e-02	0.2498	-8.400000e-03	0.1579	0.7616	0.267
dummy	Dummy Regressor	0.1958	6.270000e-02	0.2498	-8.400000e-03	0.1579	0.7616	0.013
lr	Linear Regression	342.9829	4.601515e+06	1533.7743	-7.287669e+07	1.9194	591.5804	0.295

Figura A1. Resultados de los Modelos Machine Learning propuestos para predecir CIC_log10

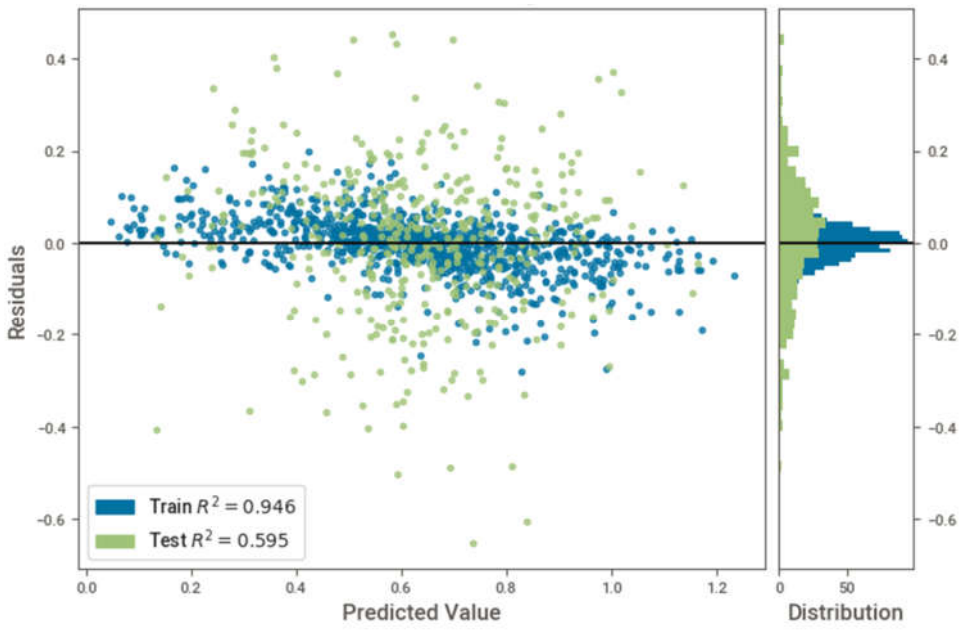

Figura A2 Residuales del modelo Random Forest Regressor