LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

PREDICCIÓN DEL DIÁMETRO PONDERADO MEDIO DE LAS PARTÍCULAS DEL SUELO EN LA VEREDA BAJO TABLAZO DEL MUNICIPIO DE MANIZALES - COLOMBIA

PREDICTING THE WEIGHTED AVERAGE DIAMETER OF SOIL PARTICLES IN THE BAJO TABLAZO AREA OF THE MUNICIPALITY OF MANIZALES - COLOMBIA

Alex Junior Matta Morales, ajmattam@libertadores.edu.co, José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co, Fundación Universitaria Los Libertadores

RESUMEN

La estructura del suelo es el arreglo y la organización de las partículas que lo constituyen, su importancia radica en que permite establecer relaciones con la erosión en sistemas de conservación de suelos. Su evaluación en campo desde las características de tipo y clase, se complementan con la medición del grado de desarrollo que presenta, sin embargo, por causa de la cantidad y costo de las muestras, se hace necesario recurrir a métodos predictivos para su determinación y establecer una función de edafotransferencia. Se evaluaron cuatro modelos de regresión basados en aprendizaje supervisado de machine learning: Regresión Ridge, Regresión Lineal, Regresión Lasso y Árbol de Decisiones; para predecir el tamaño de partículas estructurales del suelo (Diámetro Ponderado medio en mm) a partir de propiedades obtenidas en campo junto

con análisis químicos de suelos en la zona agrícola de la vereda Bajo Tablazo del municipio de Manizales - Colombia. Mediante el entrenamiento del conjunto de datos La Floresta con 15 variables entre físicas, químicas y biológicas; las métricas RMSE, MSE y R^2 permitieron evaluar el desempeño de los modelos. El modelo de predicción entrenado se aplicó al conjunto de datos de la vereda Bajo Tablazo para obtener una predicción del DPM_mm. El valor de R^2 indicó que los modelos no están explicando la varianza de la variable objetivo a partir de las variables explicativas. A pesar que la aplicación de modelos permite la obtención de valores de propiedades del suelo, en este caso no fue posible la obtención de esta relación. Posiblemente por la baja cantidad de registros y la no heterogeneidad de los datos, es decir se requiere información más diversa de diferentes lugares para que los modelos logren generalizar el comportamiento. También es posible que una mejor preparación de las variables sea necesaria, aquí solo se aplicó Log+1 para mejorar la escala de los valores.

Palabras clave: Aprendizaje supervisado, modelos de regresión, predicción, suelo, función de edafotransferencia.

SUMMARY:

The soil structure refers to the arrangement and organization of particles that make it up. Its importance lies in establishing relationships with erosion in soil conservation systems. Field evaluation of soil structure characteristics, such as type and class, is complemented by measuring its degree of development. However, due to the quantity and cost of samples, predictive methods are necessary to determine its structure and establish an edaphotransference function. Four regression models based on supervised machine learning were evaluated: Ridge Regression, Linear Regression, Lasso Regression, and Decision Tree. They were used to predict the size of structural soil particles (Weighted Mean Diameter in mm) from field properties and soil chemical analysis in the agricultural

area of Bajo Tablazo, Manizales - Colombia. The models' performance was evaluated using RMSE, MSE, and R2 metrics after training the La Floresta dataset with 15 physical, chemical, and biological variables. The trained prediction model was applied to the Bajo Tablazo dataset to obtain a DPM_mm prediction. However, the R2 value indicated that the models were not explaining the variance of the target variable from the explanatory variables. Although models can provide soil property values, it was not possible in this case due to the low number of records and non-heterogeneous data. More diverse information from different locations is required for the models to generalize behavior better. Additionally, a better preparation of the variables may be necessary, as only Log+1 was applied to improve the scale of the values.

Keywords: Supervised learning, regression models, prediction, soil, edaphotransference function.

INTRODUCCIÓN

Para la evaluación y medición de propiedades del suelo se requiere cierta cantidad de muestreos en campo con el fin de abordar su distribución espacial y de este modo llegar a adecuadas decisiones de manejo, sin embargo, estas prácticas pueden tener limitaciones como disponibilidad de tiempo y costos de los análisis. Este es el caso de la evaluación cuantitativa de la estabilidad estructural por método del tamizado de Yoder (Diámetro Ponderado Medio en milímetros DPM_mm) IGAC (Montenegro y Malagón, 1990, p23). Es una prueba específica, que implica el gasto de recursos, pero, es útil en aporte técnico para trabajar con suelos de ladera y zonas lluviosas de conservación como un referente de agregación y estabilidad de los suelos. Por esta razón, se recurren a métodos predictivos, que permiten el acercamiento a esta propiedad del suelo empleando

herramientas matemáticas y estadísticas que utilizan la información mediante el entrenamiento iterativo para predecir valores de un conjunto de datos.

Las propiedades físicas, químicas o biológicas, contienen un número representativo de variables que las describen, tanto discretas como aleatorias y distribuidas en el espacio del suelo. La aplicación de modelos de predicción en propiedades del suelo se evidencia en trabajos realizados por Orellana et al (2019), en el uso de herramientas matemáticas como matrices, estadística descriptiva y regresión múltiple, permitiendo predecir la respuesta de cultivos de maíz a partir de las condiciones físicas y químicas de los suelos. Pérez et al (2017), estimaron valores edáficos mediante el empleo de técnicas de modelización espacial, basada en procesos y variables ambientales, para la predicción de la materia orgánica del suelo. Valderrama et al (2021, p11), desarrollaron un modelo predictivo de propiedades mecánicas del suelo usando redes neuronales para el diseño de pavimentos en obras civiles. Castaño y González (2022), emplearon modelos supervisados de machine learning con el fin de predecir y establecer las variables que influyen sobre la CIC en el cultivo de aguacate. En general se pueden entrenar modelos para la predicción de valores de otras variables o propiedades del suelo con ciertas limitaciones en similares condiciones agrológicas.

Según Baver, citado por el Montenegro (p59) la estructura es una propiedad física considerada como la clave de la productividad del suelo, sin desconocer que existe una estrecha relación entre las propiedades físicas y químicas, está asociada con la acción biológica, el fenómeno de expansión y contracción de arcillas, el pH y la morfología … favorece los procesos de agregación e intercambio gaseoso, la retención de humedad y un efecto modificante en la textura …. es decir, es el arreglo y la organización de las partículas constitutivas del suelo Montenegro (p244). Para interpretar la estabilidad estructural con relación al DPM_mm (IGAC, p37), se emplean los rangos de la tabla 1:

Tabla 1. Rangos de interpretación de la estabilidad estructural en mm y log_{10}+1.

DPM mm	DPM log10+1	Interpretación
< 0.5	< 0.6989	Inestable
0.5 – 1.5	0.6989 - 1.176	Ligeramente estable
1.5 - 3	1.176 - 1.4771	Moderadamente estable
3 - 5	1.4771 - 1.6989	Estable
> 5	1.6989	Muy estable

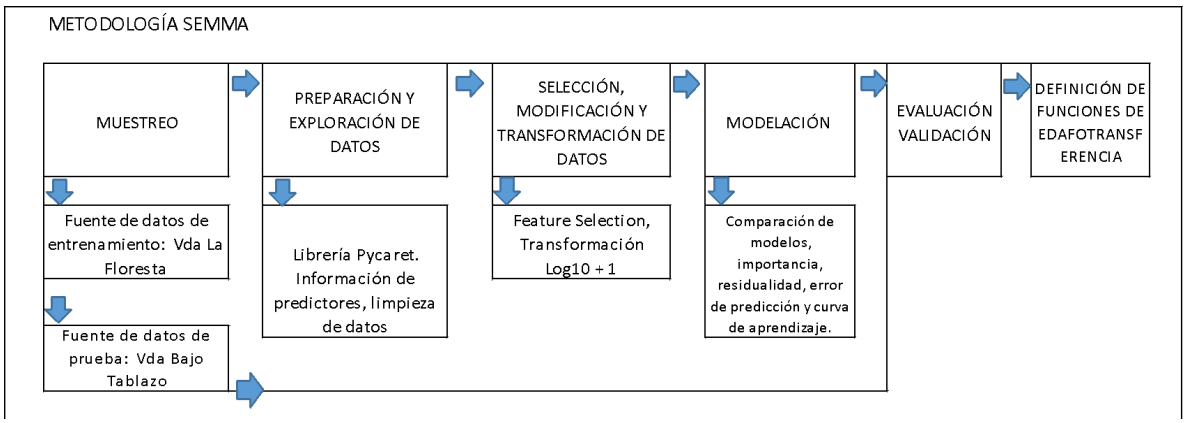
El Anexo 1 área de estudio, muestra que las dos zonas muestreadas poseen las mismas condiciones agrológicas y de uso, que corresponden a un paisaje de vertiente de montaña, de clima húmedo, en ladera, suelo derivado de ceniza volcánica, muy profundo, rico en materia orgánica, bien drenado y aireado, de color negro a pardo muy oscuro, textura franco arenosa, ligeramente ácido, contenido de calcio, magnesio y potasio bajo y disponibilidad baja de fósforo asimilable por la planta, perteneciente a la asociación de suelos Chinchiná - Azufrado Typic Dystrandepts (mapa suelos caldas 100K, IGAC).

Como hipótesis experimental se tiene que, a partir del entrenamiento de información sobre propiedades del suelo en modelos de predicción supervisada, se podrá dentro de un margen de rendimiento y error, predecir los valores de diámetro ponderado medio de las partículas del suelo DPM_mm. Por tanto, el objetivo de investigación es: aplicar cuatro modelos de predicción supervisada, para predecir la propiedad física de estabilidad estructural representada como el Diámetro Ponderado Medio DPM_mm de las partículas del suelo para predios de producción agrícola de la vereda Bajo Tablazo de Manizales - Colombia.

METODOLOGÍA

Para la predicción de la estabilidad estructural (DPM_mm) se implementó aprendizaje supervisado con cuatro algoritmos de regresión: Regresor de Ridge ($ridge$), Regresión Lineal (lr), Regresor de Lasso ($lasso$) y el Regresor de Árbol de Decisiones (dt). La selección de estos modelos es por la baja cantidad de registros disponibles. El proceso se desarrolló mediante la metodología SEMMA de SAS Instituto (2023), en las fases de muestreo, exploración de datos, modificación, modelación y evaluación, como se muestra en la figura 1.

Figura 1. Flujo de trabajo empleando la metodología SEMMA.



Muestreo

La fuente de datos empleada para el entrenamiento y creación de los modelos se dividió en 90% para el entrenamiento y 10% para validación, ésta se obtuvo de los resultados de investigación realizados por Matta y Ramírez (2006) en Indicadores de calidad del suelo y variabilidad espacial. Los datos corresponden a una matriz inicial de 76 filas x 26 variables obtenidas de análisis químico de suelos del IGAC, pruebas físicas del laboratorio de suelos de la facultad de agronomía de la Universidad de Caldas y muestras biológicas tomadas en el predio La Celia de la vereda La Floresta en Villamaría Caldas Colombia. El conjunto de datos de aplicación corresponde a una matriz de 41 filas x 19 variables que contiene los análisis químicos de suelos realizados en laboratorio IGAC en 2007, provenientes de 31 predios de la vereda Bajo Tablazo del municipio de Manizales -

Colombia, los resultados son propiedad de los agricultores y obtenidos mediante reunión con los mismos.

Exploración de datos y modificación

Los conjuntos de datos de entrenamiento y aplicación se organizaron de manera que correspondan con el mismo número y etiqueta de variables cuantitativas continuas. La preparación inició con la verificación de valores nulos y variables incompletas del conjunto de entrenamiento. Se evaluó la distribución de las variables y aquellas que no presentaron normalidad se transformaron empleando la función logaritmo (log10+1). La variable objetivo es el Logaritmo del Diámetro Ponderado Medio de las partículas del suelo en milímetros (DPM_mm_log10) y las variables predictoras son densidad aparente (gr/cm^2), población de lombrices (ind/m^2), longitud radicular (cm), pH, carbono orgánico (%), CIC (me/100gr), magnesio, potasio, sodio, calcio, bases totales, fósforo (ppm), arena, limo y arcilla (%). Definiendo una matriz inicial de 76 instancias por 15 vectores columna.

Modelación y evaluación

Se incluyó la selección de características controlando la baja varianza y multicolinealidad; permitiendo que del conjunto de datos se excluyan aquellos predictores redundantes o que no aportan a la predictibilidad. Se ejecutó el algoritmo de comparación de modelos para verificar las medidas de rendimiento de los regresores con los predictores seleccionados. Tanto el RMSE como el MSE son formas de medir la distancia entre dos vectores: el vector de predicciones y el vector de valores objetivo (Geron, 2019, p74). La MSE es la suma de los cuadrados de las diferencias entre los valores predichos por el modelo y los valores reales, y luego dividiendo la suma por el número de observaciones. Como expresa la ecuación 1, Geron (2019, p169):

$$MSE = (1/n) * SUM((Y_{PRED} - Y_{TRUE})^2) \quad \textbf{(1)}$$

Donde: n es el número de observaciones, y_pred son los valores predichos por el modelo e y_true son los valores reales. La RMSE se calcula a partir de la ecuación 2:

$$RMSE = sqrt(MSE) \quad \textbf{(2)}$$

A cada regresor se le aplicó el algoritmo de evaluación de modelo para obtener el error de predicción, residuales y la curva de aprendizaje. Para la prueba de evaluación se utilizó el conjunto de datos Bajo tablazo. Se transformaron previamente a (Log10+1) y se ejecutó el algoritmo de predicción con cada uno de los modelos, obteniendo los valores del DPM_mm_log10 de las partículas del suelo. Se realizó la descripción comparativa de las métricas de los regresores y se emitieron conclusiones.

RESULTADOS

Los predictores con mejores características seleccionados por el control de multicolinealidad y baja varianza de los datos de entrenamiento fueron la población de lombrices por m^2 (Lombmsq_log10), % carbono orgánico (Pcarb_org_log10), fósforo (Fosforo_log10), % arena (Arena_log10) y % limo (Limo_log10).

La tabla 2, muestra que el DPM_mm_log10 presenta una relación débil a moderada con los predictores, desde la magnitud y dirección positiva el % carbono orgánico obtuvo coeficiente Pearson(P) de 0.39, mientras que con población de lombrices, arena y limo la correlación se presentó en con valores cercanos cero. El fósforo presentó una correlación negativa (P= -0.28).

Tabla 2. Matriz de correlación de Pearson del conjunto de datos de entrenamiento La Floresta.

Objetivo\Predictor	Lombmsq_log10	Pcarb_org_log10	Fosforo_log10	Arena_log10	Limo_log10
DPM_mm_log10	0.166761	0.385560	-0.283782	0.097412	0.021168

En la comparación de modelos, se evidenció que la regresión de ridge y lineal presentaron los valores de MSE y RMSE más bajos indicando que ambos tienen ajuste a los datos y que las predicciones están cercanas a los valores reales. El regresor dt y lasso presentaron sobreajuste y sub-ajuste respectivamente. En la tabla 3, se muestran las métricas de los cuatro regresores. Desde la capacidad predictiva el rendimiento en datos no vistos, el coeficiente de determinación R^2 fue negativo para el modelo de regresión lineal, indicando que no explica adecuadamente la varianza de la variable objetivo DPM_mm_log10, por tanto, aunque hay ajuste de datos y que las predicciones son cercanas a valores reales, no hay una adecuada varianza explicada en relación al DPM_mm_log10.

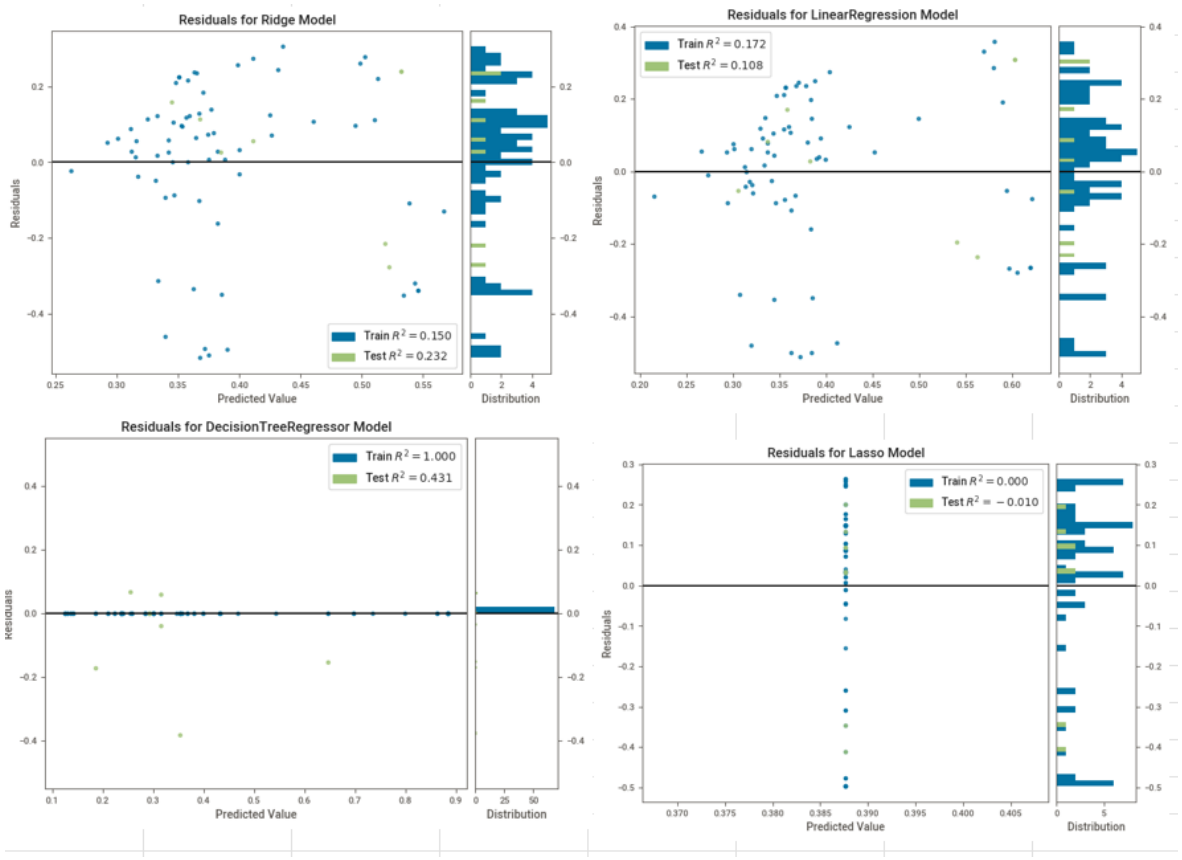
Tabla 3. Métricas de cuatro modelos de regresión en validación cruzada, por ejecución del algoritmo comparación de modelos.

	Modelo	MSE	RMSE	R2
ridge	**Ridge Regression**	**0.0512**	**0.2202**	**-0.4392**
lr	**Linear Regression**	**0.0517**	**0.2225**	**-0.5576**
dt	**Decision Tree Regressor**	**0.0667**	**0.2522**	**-1.2136**
lasso	**Lasso Regression**	**0.0564**	**0.2271**	**-0.4392**

En la figura 4, los residuales de los modelos de ridge y lineal muestran que la diferencia entre el R^2train y R^2test, fue entre 0,082 y 0,064 unidades absolutas respectivamente, esta

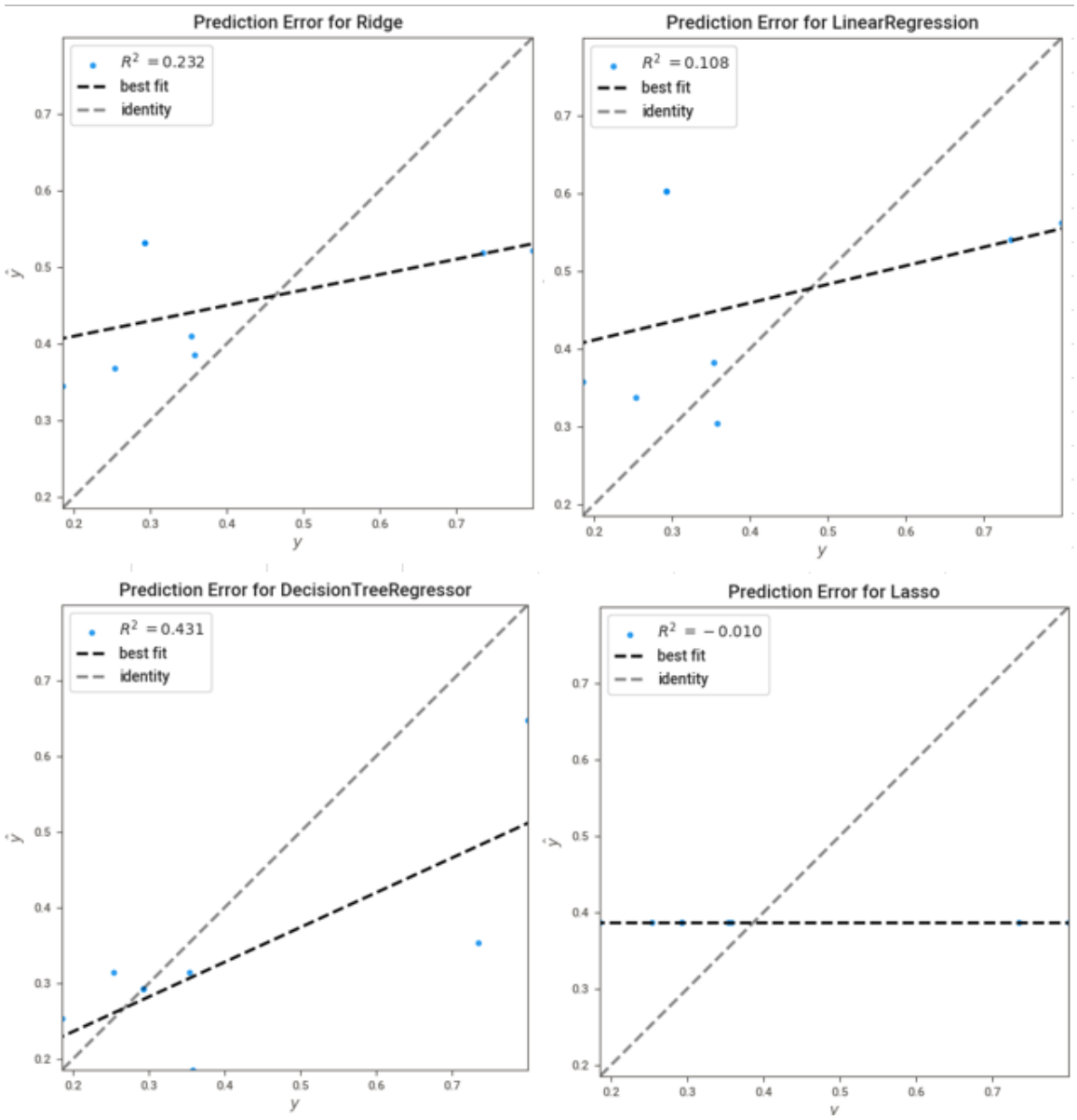
corta diferencia, aunque es buen indicador, está acompañado de la presencia de puntos alejados de la línea cero, indicando que los modelos están presentando problemas de ajuste. Además, se observa que los modelos dt y lasso están sobre-ajustados.

Figura 4. Residuales en cuatro modelos de regresión.



En la figura 5 desde la calidad predictiva de los modelos, se observa que el error de predicción para lineal y ridge mostró un R^2 con valores positivos de 0.108 y 0.232 respectivamente, indicando que hay una relación débil entre los valores predichos y los valores reales; ambos modelos tienen sesgo y no están capturando adecuadamente la relación entre las variables. Para dt se evidenció que la relación entre variables fue positiva R^2 =0.43 indicando menor varianza explicada que ridge y lr.

Figura 5. Gráficas de error de predicción de cuatro modelos de regresión en los datos de testeo.



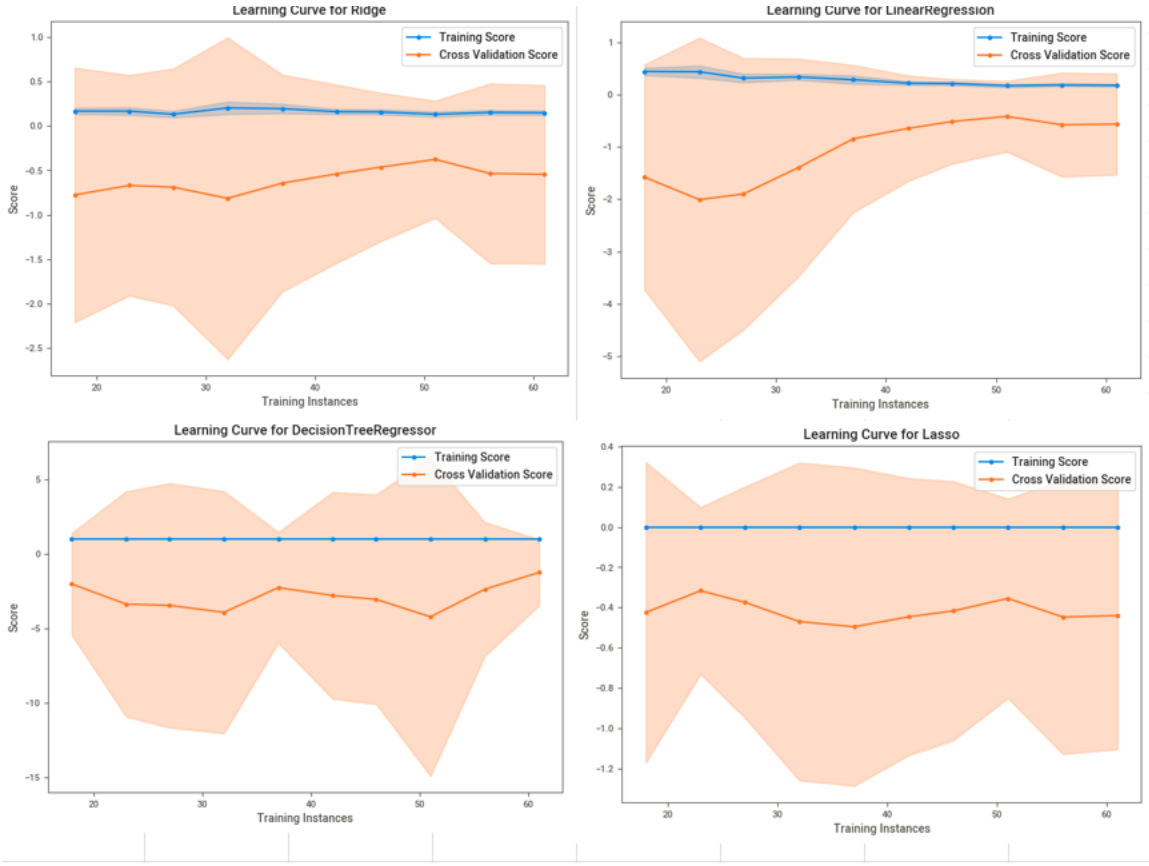
En la tabla 4, se observan los resultados de las métricas al ejecutar la predicción con los datos de la vereda bajo tablazo. Lineal y lasso presentaron rendimientos que evidencian un bajo ajuste de los datos y que las predicciones pueden estar alejadas de los valores reales, pues presentan los más altos RMSE. Ridge y dt presentaron valores del RMSE de 0.18 y 0.16 con distancias entre valores reales y predichos es de 0.02 y 0.03 unidades logarítmicas.

Tabla 4. Métricas de las predicciones en cada modelo de regresión para los datos de validación.

Model	MSE	RMSE	R2
Linear Regression	0.0407	0.2017	0.1080
Decision Tree Regressor	0.0259	0.1611	
Ridge Regression	0.0350	0.1871	
Lasso Regression	0.0460	0.2146	

En la figura 6, las curvas de aprendizaje muestran la evolución del rendimiento de los modelos para aprender a medida que aumenta la cantidad de datos (Geron, 2019, p191). Los modelos de ridge, lineal y dt, presentaron menor sobreajuste, así como una tendencia a la estabilidad a partir de 50 instancias en adelante.

Figura 6. Curvas de aprendizaje de cuatro modelos de regresión.



DISCUSIÓN DE RESULTADOS

Los predictores seleccionados presentaron relación con el DPM_mm_log10, el % carbono orgánico en el suelo influencia la agregación de las partículas, haciéndolo más estable, este resultado coincide con lo expuesto por Montenegro (p622), donde las adiciones de materia orgánica favorecen la formación de la estructura. Así mismo, % de Arena y % Limo son predictores inherentes de la textura del suelo, y por tanto según Montenegro (p35), ambas propiedades físicas deben interpretarse conjuntamente debido a las modificaciones que la estructura induce en la textura.

De acuerdo a la metodología planteada y los resultados obtenidos, en la evaluación de los cuatro modelos para la predicción de DPM_mm_log10, se encontraron problemas de bajo ajuste y sobre-ajuste entre los valores reales y los valores predichos, como se indica en la tabla 3.

En la tabla 5 se muestra el alcance de cada regresor, el modelo Ridge y lineal presentaron medias por encima de la media de los datos de origen, indicando sobreestimación del diámetro ponderado medio. El modelo lasso simplificó las predicciones al mismo valor de la media de origen y con desviación cero. Del mismo modo dt presentó medias inferiores al origen, indicando subestimación de la variable objetivo. La alta desviación que presentó la variable objetivo original para el entrenamiento evidenció la dificultad en la regresión lineal para obtener valores de R^2 positivos.

Tabla 5. Media del DPM_mm_log10 en los momentos de entrenamiento y validación para cada modelo de regresión y las medias de la variable objetivo de origen.

DPM_log10		**Original**	**lr**	**dt**	**ridge**	**lasso**
Entrenamiento La Floresta	mean	0,389	0,461	0,332	0,452	0,387
	std	0,232	0,127	0,137	0,082	0

validación bajo tablazo	mean	0,495	0,359	0,479	0,387
	std	0,054	0,152	0,056	0

La tabla 2, de correlaciones con el DPM_mm_log10 muestra una moderada correlación con la población de lombrices y el carbono orgánico principalmente, lo que permite la identificación de funciones de edafotransferencia. Según Izquierdo (2020), corresponden a funciones continuas, simples o totalmente empíricas: puesto que se hacen relaciones entre las variables predictoras y las estimadas al usar un análisis de regresión o modelos matemáticos. Es decir, que cada modelo de regresión desarrollado es una función de edafotransferencia con relaciones entre las variables que la componen.

La función de edafotransferencia establecida por el regresor lineal en la ecuación 3, es una función que calcula el DPM_mm_log10, para condiciones agrológicas similares a las áreas de estudio, sin embargo la función presenta sobre ajuste y no explica adecuadamente la varianza.

$$DPMmmlog = -0.65344036 + 0.09146022(Pcarborlog10) + 0.47734392(Arenalog10) - 0.03004846(Limolog10) + 0.35367092(Lombmsqlog10) + 0.17674652(Fosforolog10) \quad (3)$$

CONCLUSIONES

El valor R^2 indica que los modelos no están explicando la varianza de la variable objetivo a partir de los predictores, aunque hayan presentado cierto grado de correlación. Una razón que explica el bajo ajuste, es porque el conjunto de datos de entrenamiento solo contaba con 76 instancias y según un artículo de Hastie, Tibshirani y Friedman (2009), la precisión de los modelos de regresión puede aumentar significativamente con un aumento en el

tamaño de los datos de entrenamiento. Al igual que Bengio, Courville y Vincent (2013) mencionan que un mayor número de instancias de entrenamiento puede ayudar al modelo a generalizar mejor, ya que el modelo puede aprender una representación más precisa de la distribución de los datos.

REFERENCIAS BIBLIOGRÁFICAS

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

Castaño, M. H. (2022). Predicción de la capacidad de intercambio catiónico (CIC) en cultivos de aguacate empleando modelos machine learning. Recuperado de: http://hdl.handle.net/11371/4736.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media. ISBN: 978-1-492-03264-9.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer Science & Business Media.

Montenegro, H., & Malagón, D. (1990). Propiedades físicas de los suelos. Instituto Geográfico Agustín Codazzi.

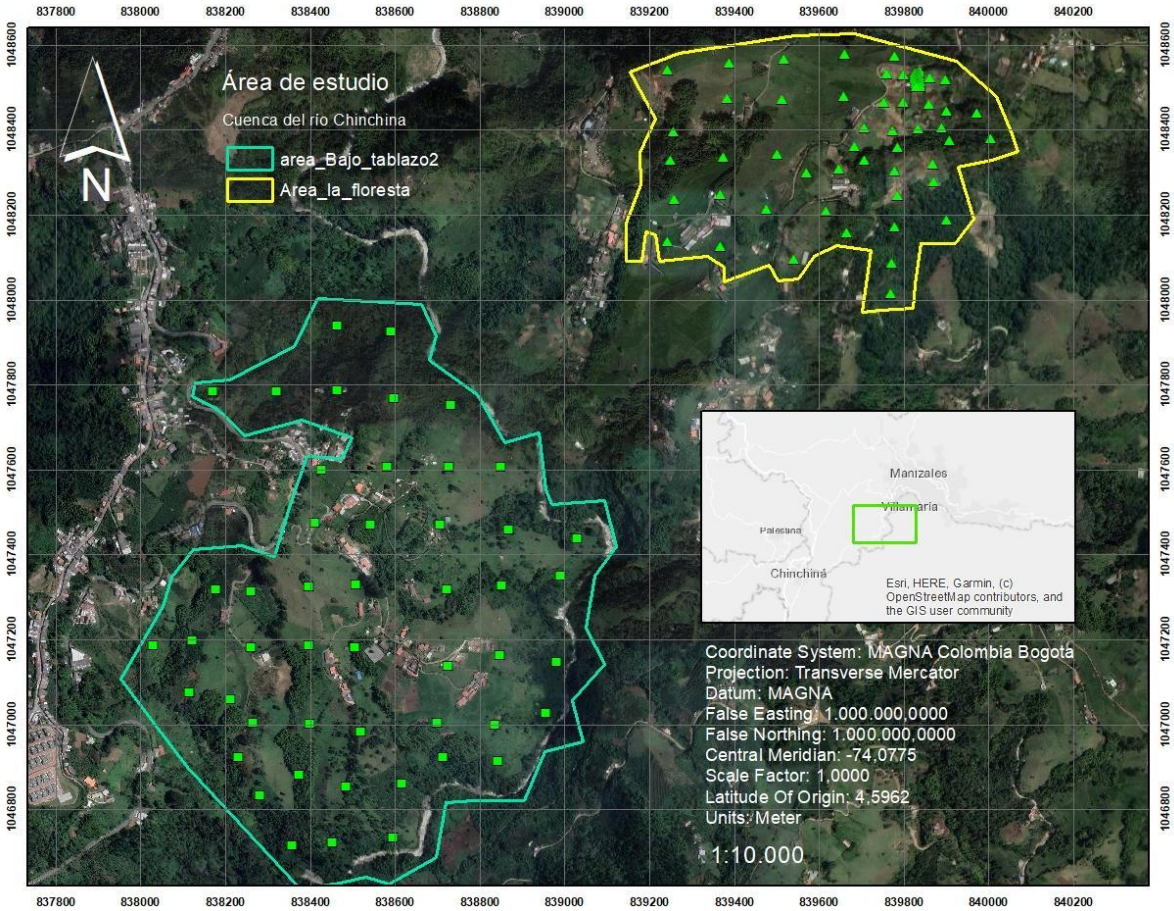
Orellana, R., Párraga, C., & Uzcátegui, J. M. (2019). Análisis predictivo de la respuesta del maíz (Zea mays L.) a las condiciones físicas y químicas de los suelos agrícolas, municipio Guanare, estado Portuguesa, Venezuela. Ciencia y Tecnología Agropecuaria, 4(2), 63-68.

Pérez Cutillas, P., Barberá, G. G., & Conesa García, C. (n.d.). Efectos de las variables ambientales en la estimación de materia orgánica del suelo a escala regional en un ambiente semiárido (región de Murcia, España). Departamento de Geografía, Universidad de Murcia. Grupo de Erosión y Conservación de Suelos, Centro de Edafología y Biología Aplicada del Segura, Consejo Superior de Investigaciones Científicas (CEBAS-CSIC).

Valderrama-Purizaca, F. J., Mejía-Cabrera, H. I., Muñoz Pérez, S. P., & Tuesta-Monteza, V. A. (2021). Desarrollo de un modelo predictivo de las propiedades mecánicas del suelo usando redes neuronales artificiales. Ciencia y Tecnología Agropecuaria, 22(4), 1-15. doi: https://doi.org/10.21501/21454086.4042

ANEXOS

Anexo 1. Área de estudio, municipios de Manizales y Villamaría - Caldas - Colombia



Fuente: Realizada por el autor, en QGIS.