PREDICCIÓN DE CONTENIDO DE ALUMINIO Y NIVELES DE PH EN SUELO AGRÍCOLA

Manuel Alejandro González, Santiago Alonso

INTRODUCCIÓN

La medición de los niveles de materia orgánica e inorgánica y pH en el suelo son fundamentales para el éxito de la agricultura, ya que conociéndolos se pueden tomar decisiones informadas y estratégicas para optimizar el uso de recursos y mejorar la salud de los cultivos. Una de las variables más importantes es el nivel de Aluminio en el suelo, ya que este este es uno de los factores limitantes en el crecimiento de los cultivos [2]. En Colombia, al menos el 50% de los suelos están afectados por problemas de toxicidad debido al aluminio [2].

La fácil medición de estas variables beneficiaría entonces al agricultor colombiano. Para obtener información acerca de los suelos, se deben someter muestras de este a distintas pruebas complicadas y costosas. Una alternativa, sería realizar una espectrometría con la muestra, sin embargo, esto aun no permite conocer todas las variables necesarias para determinar el estado del suelo.

En este proyecto, se utilizaron datos de muestras de suelo de cultivos en diferentes regiones de Colombia, junto con las mediciones de radiancia en diferentes espectros de luz, y su nivel de aluminio y pH del suelo medidos con técnicas convencionales. Utilizando estos datos, se implementaron modelos de regresión por vectores de soporte y redes neuronales para predecir los niveles de pH y aluminio de las muestras.

Los resultados de este proyecto tienen el potencial de proporcionar fácilmente a los agricultores y profesionales del sector agrícola una herramienta efectiva y rápida para la toma de decisiones basada en datos, lo cual puede contribuir a la optimización de la producción agrícola, la mejora de la calidad de los cultivos.

METODOLOGÍA

El data set trabajado fue limpiado de valores no numéricos y desconocidos. Luego se utilizaron dos métodos para predecir tanto el nivel de pH como el de aluminio. El primer método fue el de redes neuronales (*Multi-layer Perceptron regressor)*. El segundo método fue utilizando vectores de soporte (*Epsilon-Support Vector Regression*). Ambos modelos fueron hechos usando la librería SkLearn. Para lograr el mejor método, se evaluó el coeficiente de determinación de Cross-validaciones para diferentes aspectos del modelo: como la normalización, la reducción dimensional (usando PCA) y los diferentes hiperparámetros de los métodos. Una vez se encontró el modelo con el mejor coeficiente de determinación, este se evaluó sobre un conjunto de prueba, el cual contenía el 10% de los datos iniciales.

RESULTADOS

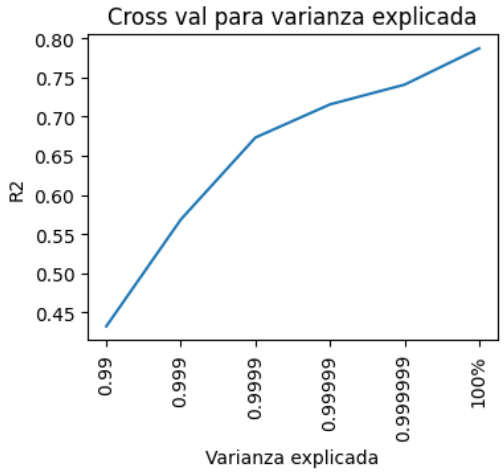
Resultados limpieza de los datos

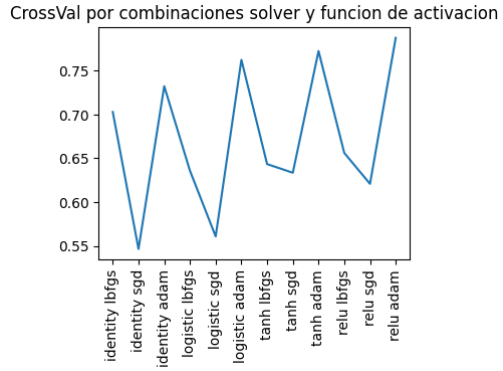
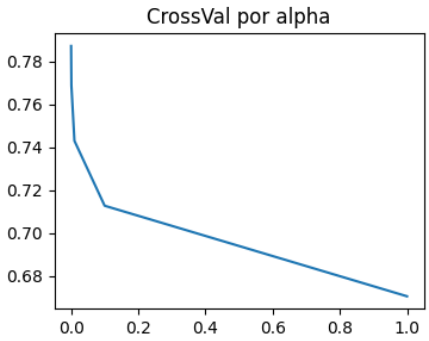
El data set inicialmente contaba con 6178 muestras. Sin embargo, muchas de estas tenían valores NA. Se decidió eliminar las muestras con alguno de estos NA porque obteníamos de todas formas una cantidad grande de datos. Para los datos de la regresión de aluminio, se trabajó con 1599 muestras finales. Para la regresión de pH, se trabajó con 3178 muestras finales

Resultados modelo de redes neuronales para la predicción de pH:

Modelo con…	**Coeficiente de determinación**
Datos sin normalizar (hiperparámetros por defecto)	0.457
Datos normalizados (hiperparámetros por defecto)	0.561
Datos normalizados con PCA (100% de los componentes) (hiperparámetros por defecto)	0.787

Para concluir que la mejor PCA era aquella que utilizaba todos los componentes, se realizó una Cross-validación para diferentes valores de varianza explicada:



Iteración de Hiperparámetros	Gráficas de la iteración
Combinaciones de las diferentes funciones de activación y optimizador	
Diferentes valores para Alpha (taza de aprendizaje)	

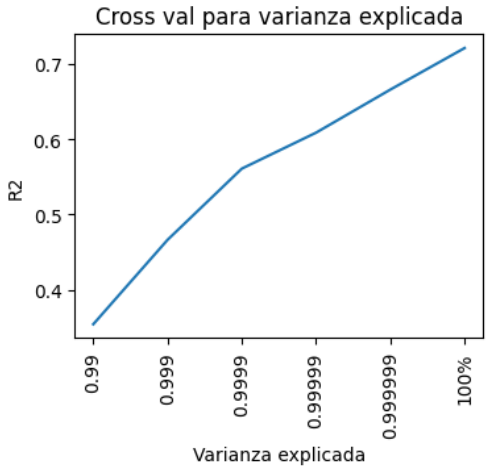
Según los resultados anteriores, el modelo final se ajustó con los datos normalizados, transformados con una PCA (donde se dejó el 100% de los componentes) y se usó la red neuronal con función de activación ReLU, optimizador Adam y taza de aprendizaje por defecto.

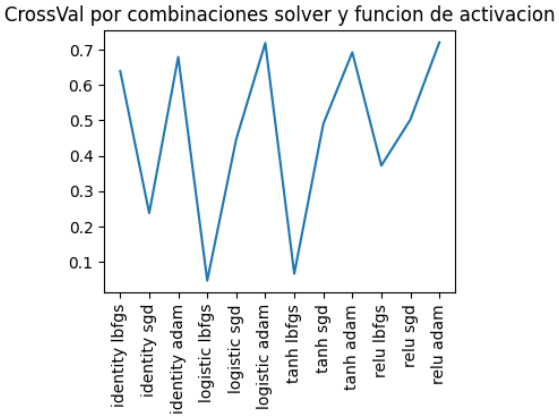
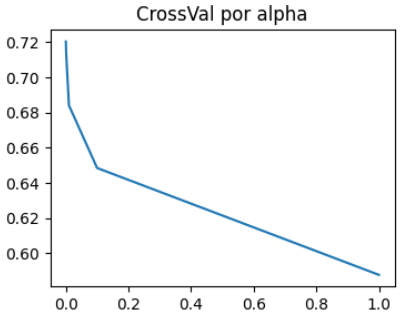
Al evaluar el modelo final con los datos de prueba (10% de los datos) se obtuvo un coeficiente de determinación de 0.7685.

Resultados modelo de redes neuronales para la predicción de Aluminio:

Modelo con…	Coeficiente de determinación
Datos sin normalizar (hiperparámetros por defecto)	0.205
Datos normalizados (hiperparámetros por defecto)	0.477
Datos normalizados con PCA (100% de los componentes) (hiperparámetros por defecto)	0.720

Para concluir que la mejor PCA era aquella que utilizaba todos los componentes, se realizó una Cross-validación para diferentes valores de varianza explicada:



Iteración de Hiperparámetros	Gráficas de la iteración
Combinaciones de las diferentes funciones de activación y optimizador	
Diferentes valores para Alpha (taza de aprendizaje)	

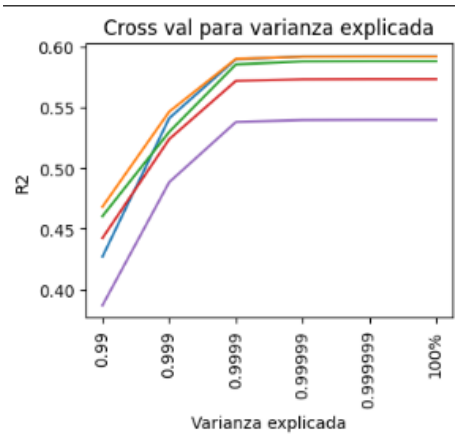
Según los resultados anteriores, el modelo final se ajustó con los datos normalizados, transformados con una PCA (donde se dejó el 100% de los componentes) y se usó la red neuronal con función de activación ReLU, optimizador Adam y taza de aprendizaje por defecto.

Al evaluar el modelo final con los datos de prueba (10% de los datos) se obtuvo un coeficiente de determinación de 0.6568.

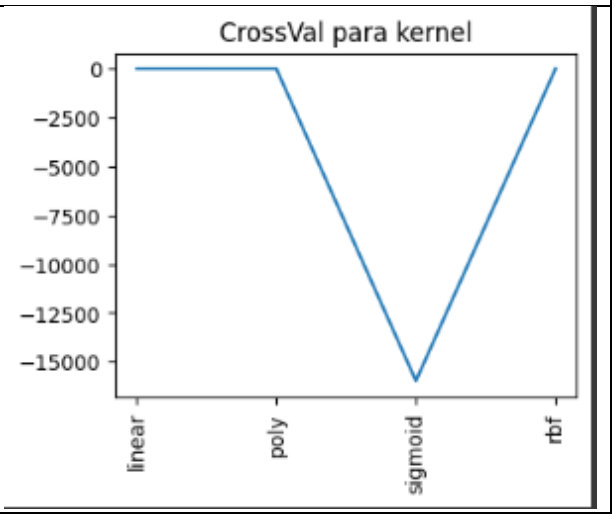
Resultados modelo de SVM para la predicción de pH:

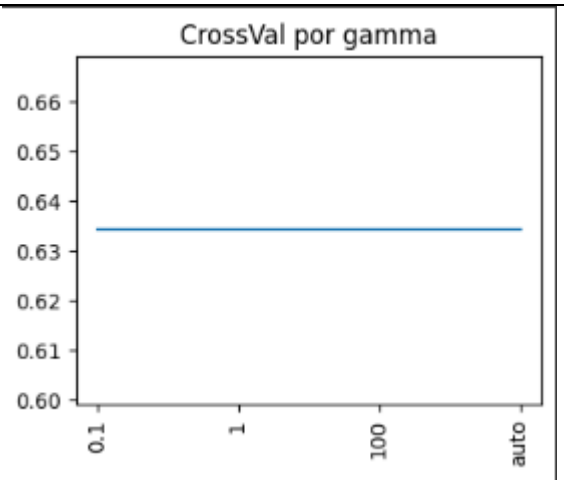
Modelo con…	Coeficiente de determinación
Datos sin normalizar (hiperparámetros por defecto)	0.532
Datos normalizados (hiperparámetros por defecto)	0.576
Datos normalizados con PCA (100% de los componentes) (hiperparámetros por defecto)	0.576

Una vez obtenidos estos valores, se procedió a evaluar cual era la mejor PCA, así que se realizó una Cross-validación para diferentes valores de varianza explicada:



A partir del 0.99999 el valor de la varianza se estabiliza y llega a su máximo valor de coeficiente de determinación (R2=0.576)

Iteración de Hiperparámetros	Gráficas de la iteración
Diferentes valores para “Kernel”	

Diferentes valores para “gamma” (Influye en la capacidad de generalización y rendimiento del modelo).	

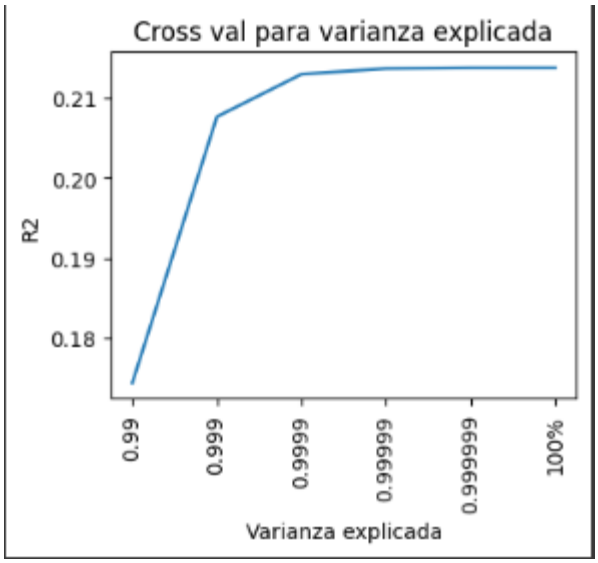
Según los resultados anteriores, el modelo final se terminó ajustando con los datos normalizados, transformados con una PCA al 100% de componentes, y se usó el SVR con un tipo de “kernel: linear” y un valor de “gamma” de 0.1.

Una vez se evaluó el modelo final con los datos de prueba (10% de los datos) se logró obtener un Coeficiente de determinación de 0.572. Esto es inferior a lo logrado mediante las redes neuronales.

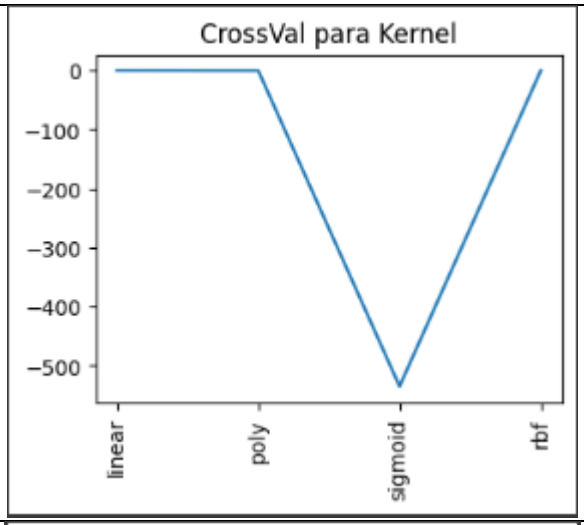
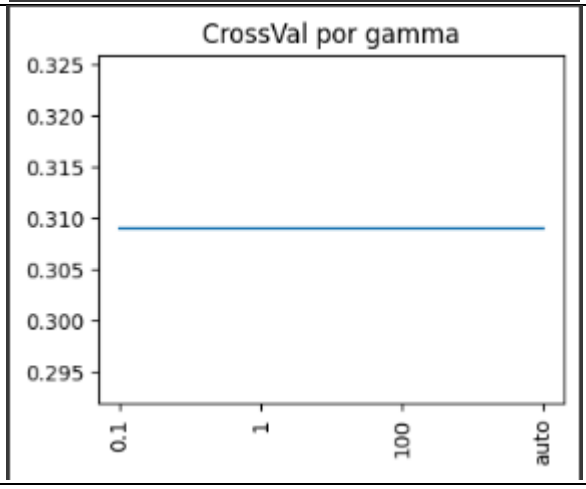
Resultados modelo de SVM para la predicción de Aluminio:

Modelo con…	**Coeficiente de determinación**
Datos sin normalizar (hiperparámetros por defecto)	0.168
Datos normalizados (hiperparámetros por defecto)	0.213
Datos normalizados con PCA (100% de los componentes) (hiperparámetros por defecto)	0.213

Una vez obtenidos estos valores, se procedió a evaluar cual era la mejor PCA, así que se realizó una Cross-validación para diferentes valores de varianza explicada:



A partir del 0.999999 el valor de la varianza se estabiliza y llega a su máximo valor de coeficiente de determinación (R2=0.213)

Iteración de Hiperparámetros	Gráficas de la iteración
Diferentes valores para “Kernel”	
Diferentes valores para “gamma” (Influye en la capacidad de generalización y rendimiento del modelo).	

Según los resultados anteriores, el modelo final se terminó ajustando con los datos normalizados, transformados con una PCA al 100% de componentes, y se usó el SVR con un tipo de “kernel: linear”, un C (parámetro que controla el equilibrio entre el ajuste de los datos de entrenamiento y la suavidad de la función de regresión.) estándar de 100 y un valor de “gamma” de “auto”.

Una vez se evaluó el modelo final con los datos de prueba (10% de los datos) se logró obtener un Coeficiente de determinación de 0.454. Esto es inferior a lo logrado mediante las redes neuronales.

CONCLUSIONES

A partir de los resultados, se encontró que los mejores modelos para predecir tanto el aluminio como el nivel de pH, son los de redes neuronales (MLP regressor), obteniendo coeficientes de determinación de 0.77 para la predicción del pH y de 0.66 para predecir aluminio. Así que para el pH, se obtuvo con las redes neuronales un modelo que logra comportarse bien y a partir de la espectrometría predecir bien esta variable, sin embargo no es perfecto y requiere de más trabajo. Para predecir aluminio, el modelo logrado no se comporta tan bien pero es un primer acercamiento para predecir este factor del suelo.

REFERENCIAS

[1] Casierra-Posada, F., & Aguilar-Avendaño, O. E. (2018). Stress for aluminum in plants: reactions in the soil, symptoms in plants and amelioration possibilities. A review. Revista Colombiana De Ciencias Hortícolas, 1(2), 246–257. https://doi.org/10.17584/rcch.2007v1i2.8701

[2] Cuervo-Álzate, Jorge Enrique, & Osorio, Nelson Walter. (2020). Gypsum incubation tests to evaluate its potential effects on acidic soils of Colombia. Revista Facultad Nacional de Agronomía Medellín, 73(3), 9349-9359. https://doi.org/10.15446/rfnam.v73n3.85259