

Modelo basado en aprendizaje de máquina para la predicción de textura de suelos

Julián Gutiérrez¹, Mariano Trillini², Federico Danilo Vallese², Railson de Oliveira Ramos³, Juan Ignacio Vanzolini⁴, Valber Elias de Almeida⁵, José Felix de Brito Neto⁶, Rodrigo Santos¹ y Marcelo Fabían Pistonesi²

¹ ICIC – CONICET, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, B8000CPB, Bahía Blanca, Argentina.

² INQUISUR – CONICET, Departamento de Química, Universidad Nacional del Sur, B8000CPB, Bahía Blanca, Argentina.

³ Departamento de Engenharia Sanitária e Ambiental, Universidade Estadual da Paraíba, Campina Grande, Paraíba 58429-500, Brazil.

⁴ Laboratorio de Suelos y Agua. INTA H. Ascasubi, Ruta 3 Km 794, 8142 Hilario Ascasubi, Argentina.

⁵ Departamento de Química, Universidade Estadual da Paraíba, Campina Grande, Paraíba 58429-500, Brazil.

⁶ Centro de Ciências Agrárias e Ambientais (CCAA), Universidade Estadual da Paraíba, Campina Grande, Paraíba 58429-500, Brazil.

julian.gutierrez@uns.edu.ar, marianotrillini94@gmail.com,
federico.vallese@uns.edu.ar, railson_uepb@outlook.com,
vanzolini.juan@inta.gob.ar, vallber_ellias@hotmail.com,
felix.brito@servidor.uepb.edu.br, ierms@criba.edu.ar,
mpistone@criba.edu.ar

Resumen En el ámbito de las ciencias del suelo, el análisis de textura del suelo es esencial, ya que proporciona información crucial sobre la composición y las propiedades físicas del mismo. Por ello, su estudio, cálculo e interpretación son fundamentales. Se propone un modelo basado en aprendizaje de máquina para predecir la textura del suelo, utilizando datos de área de píxeles de las partículas de arena, limo y arcilla obtenidos a través de imágenes digitales. Adicionalmente, datos de materia orgánica y conductividad eléctrica se integrarán al modelo como atributos. El conjunto de datos se organizará incluyendo variables objetivo como el porcentaje de arena, el porcentaje de limo y la clasificación de la textura del suelo, basada en el método de Bouyoucos, que sirve como referencia estándar de laboratorio. El desarrollo de este trabajo busca simplificar los tiempos operativos, reducir tamaño de muestra y lograr una determinación sencilla y no destructiva de la granulometría del suelo.

Palabras claves: Textura de suelos · Aprendizaje de máquina · Bosques Aleatorios · Análisis de datos.

1. Introducción

El análisis de la textura del suelo es esencial en las ciencias del suelo, proporcionando información crucial sobre su composición y propiedades físicas. Por ello, su estudio, cálculo e interpretación son fundamentales. Uno de los métodos más utilizados en esta área es el método de Bouyoucos, el cual se basa en la sedimentación y clasificación de las partículas del suelo según su tamaño. Aunque este método es ampliamente aceptado por su precisión y confiabilidad, puede ser laborioso y requerir mucho tiempo para llevar a cabo las pruebas de laboratorio [1]. En los últimos años, la aplicación de técnicas de aprendizaje automático en diversos campos científicos ha aumentado rápidamente. Esta tendencia también se ha extendido al estudio del suelo, impulsada por la abundancia de datos disponibles y la accesibilidad a algoritmos de acceso abierto. Como resultado, cada vez más investigadores están recurriendo a estas herramientas para analizar datos de suelos [4,6,7].

En un esfuerzo por mejorar la eficiencia en la determinación de la textura del suelo, se propone un modelo basado en aprendizaje de máquina. Este modelo utiliza datos de área de píxeles de las partículas de arena, limo y arcilla obtenidos a través de imágenes digitales, así como atributos adicionales como materia orgánica y conductividad eléctrica. La materia orgánica es crucial en la fertilidad del suelo y en la retención de nutrientes [5], mientras que la conductividad eléctrica puede proporcionar información sobre la salinidad y la calidad del suelo [3]. Integrar estos atributos adicionales en el modelo permite una caracterización más completa y precisa de las propiedades del suelo, lo que contribuye a mejorar la capacidad de predicción del modelo en la determinación de la textura del suelo. El objetivo de este trabajo es simplificar los tiempos operativos, reducir el tamaño de muestra y lograr una determinación sencilla y no destructiva de la granulometría del suelo.

2. Materiales y métodos

2.1. Datos de textura de suelos

Se elaboró un conjunto de datos de diferentes muestras de suelos, que comprende tanto las variables independientes como las variables dependientes para el modelo predictivo basado en aprendizaje automático. Se recolectaron 38 muestras de suelos en diversas regiones de Buenos Aires, realizando 4 réplicas a cada muestra que consistió en 500 mg de suelo en un tubo Falcon de 15 mL, llevado a volumen con agua destilada. Se efectuaron lecturas en diferentes posiciones alrededor del tubo.

En cuanto a las variables independientes, datos precisos sobre la composición del suelo se obtuvieron a partir de un sistema de imágenes especializado. Este sistema capturó el área de píxeles de la textura del suelo, permitiendo recopilar un conjunto detallado de datos de píxeles de arena y limo. Este conjunto no solo incluyó recuentos de píxeles de partículas de arena y limo, sino que también incorporó características complementarias, como el contenido de materia orgánica

y conductividad eléctrica, determinados mediante métodos estándar de laboratorio. Las variables objetivo del conjunto constan del porcentaje de arena, limo y la clasificación de la textura del suelo. Estos registros se obtuvieron mediante el método estándar de Bouyoucos que fueron brindados por el INTA (Instituto Nacional de Tecnología Agropecuaria).

Como todo análisis de datos, antes de la etapa de modelado mediante técnicas de aprendizaje automático, se realiza un análisis exploratorio de datos. De esta manera, una vez obtenido el conjunto de datos completo, se exploraron y preprocesaron los registros para comprender la estructura de los datos, identificar valores atípicos de los atributos y observar las relaciones o tendencias entre las variables.

2.2. Modelos regresores: entrenamiento, testeo y validación

Diversos modelos de regresión se emplearon para analizar el conjunto de datos y predecir los porcentajes de arena, limo y arcilla en las muestras de suelo. Estos modelos utilizan como características de entrada el recuento de píxeles de partículas de arena y limo, junto con el contenido de materia orgánica y conductividad eléctrica. Los modelos de regresión se entrenaron utilizando técnicas de aprendizaje automático adecuadas, como la regresión lineal, la regresión de vectores de soporte, la regresión bayesiana y la regresión de bosques aleatorios. Este último fue el utilizado para mostrar los resultados preliminares. A su vez, se aplicaron procedimientos de validación cruzada y ajustes de hiperparámetros para optimizar los modelos.

La eficacia de los modelos de regresión entrenados se evaluaron rigurosamente para calibrar su precisión y fiabilidad predictivas. Se calcularon parámetros como: el error cuadrático medio (ECM), el cual indica la diferencia cuadrática media entre los porcentajes de arena o limo predichos y los reales; la raíz del error cuadrático medio (RECM), que representa la desviación estándar de los residuos e indica la magnitud media de los errores de predicción; y el coeficiente de determinación (R^2) que representa la correlación entre los valores predichos y los reales.

Regresor de Bosques Aleatorios La regresión de bosque aleatorio (conocido en inglés como Random Forest Regressor) es una técnica versátil de aprendizaje automático para predecir valores numéricos. Combina las predicciones de múltiples árboles de decisión para reducir el sobreajuste y mejorar la precisión [2]. Además, se utiliza la técnica de "out of bag" (OOB), que permite evaluar el modelo de manera eficiente sin necesidad de una partición adicional de validación de los datos. Se calcula utilizando las muestras que no se utilizan en el entrenamiento del modelo, lo que se denomina muestras fuera de la bolsa. Estas muestras se utilizan para proporcionar una estimación no sesgada del rendimiento del modelo, que se conoce como puntuación OOB.

3. Resultados preliminares

En este estudio, se entrenaron y validaron modelos de regresión de aprendizaje automático utilizando un conjunto de datos diverso que incluye datos a nivel de píxel extraídos de imágenes de suelos, complementados con características adicionales como el contenido de materia orgánica y conductividad eléctrica. Mediante análisis exploratorio de los datos y resultados de parámetros de error, evaluamos la eficacia de los modelos para predecir con precisión los porcentajes de arena, limo y arcilla.

Se evaluó el rendimiento del modelo regresor de bosques aleatorios en la predicción del contenido de arena y limo para los datos de testeo. El modelo muestra un ECM de 0,728 y 0,200 para la predicción de arena y limo, respectivamente. El RECM fue de 0,853 y 0,448. Por último, se determinó que el coeficiente de determinación (R^2) fue de 0,999 tanto para la arena como el limo, lo que sugiere una fuerte correlación entre los valores de porcentajes predichos y los valores reales que se obtuvieron mediante Bouyoucos.

4. Trabajo a futuro

En esta primer instancia, se desarrolló un método innovador para predecir el porcentaje de textura del suelo (arena, limo y arcilla), en donde el modelo regresor fue entrenado para relacionar los atributos del conjuntos de datos con las variables objetivo. Como se mencionó anteriormente, estas variables dependientes, como el porcentaje de arena y limo, fueron obtenidos a partir del método tradicional de Bouyoucos. Ahora, como trabajo a desarrollar a futuro, tiene como objetivo principal relacionar las características del conjunto de datos (píxeles de arena, píxeles de limo, materia orgánica y conductividad eléctrica) con datos de porcentajes de arena, limo y arcilla obtenidos en muestras de suelos mediante el método de la pipeta de Robinson, el cual obtiene resultados más precisos. Esto mejorará aún más la predicción del modelo.

Referencias

1. Bouyoucos, G.J.: Hydrometer method improved for making particle size analyses of soils¹. *Agronomy Journal* **54**(5), 464–465 (1962). <https://doi.org/10.2134/agronj1962.00021962005400050028x>
2. Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (Oct 2001). <https://doi.org/10.1023/A:1010933404324>
3. Cremona, M.V., Enriquez, Soledad, A.: Algunas propiedades del suelo que condicionan su comportamiento: El ph y la conductividad eléctrica. *Instituto Nacional de Tecnología Agropecuaria* **73**(5-8) (7-2020)
4. Ghabi, M.: Classification of soil texture using machine learning technique. In: 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA). pp. 1–4 (2023). <https://doi.org/10.1109/INISTA59065.2023.10310648>
5. I: Organic Matter and Soil Quality, pp. 113–135. Springer Netherlands, Dordrecht (2008). https://doi.org/10.1007/978-1-4020-8433-1_4

6. Maruthaiah, T., Vajravelu, S.K., Kaliyaperumal, V., Kalaivanan, D.: Soil texture identification using libs data combined with machine learning algorithm. *Optik* **278**, 170691 (2023). <https://doi.org/10.1016/j.ijleo.2023.170691>
7. Mirzaeitalarposhti, R., Shafizadeh-Moghadam, H., Taghizadeh-Mehrjardi, R., Demyan, M.S.: Digital soil texture mapping and spatial transferability of machine learning models using sentinel-1, sentinel-2, and terrain-derived covariates. *Remote Sensing* **14**(23) (2022). <https://doi.org/10.3390/rs14235909>