ISSN 1870-4069

Estudio de técnicas de aprendizaje automático para la estimación de la humedad del suelo en agricultura

Noel A. Zavala-Díaz, Juan C. Olivares-Rojas,
Jonathan Zavala-Díaz, Enrique Reyes-Archundia,
Adriana Téllez-Anguiano, Gerardo M. Chávez-Campos,
Arturo Méndez-Patiño

Tecnológico Nacional de México
División de Estudios de Posgrado e Investigación
México

{m21121662, juan.or, d19123006, enrique.ra, adriana.ta,
gerardo.cc, arturo.mp}@morelia.tecnm.mx

Resumen. La humedad del suelo es crucial en diversos campos, y su monitoreo para guiar el riego representa un desafío. El aprendizaje automático ha surgido como una herramienta prometedora para predecir con precisión los niveles de humedad del suelo. Este estudio se centra en evaluar técnicas de aprendizaje automático para esta tarea, entrenando modelos con variables meteorológicas y mediciones directas de humedad del suelo. Se implementaron cuatro algoritmos de aprendizaje automático, destacando el Gradient Boosting Regressor como el más efectivo. Además, se presenta un conjunto de datos procesado que combina mediciones meteorológicas y de humedad del suelo, esperando que sea útil para futuras investigaciones. Este enfoque busca mejorar la comprensión y la capacidad de previsión de la humedad del suelo, crucial para la planificación agrícola y la gestión del agua en la agricultura.

Palabras clave: Humedad del suelo, aprendizaje automático, modelos de regresión.

Study of Machine Learning Techniques for the Estimation of Soil Moisture in Agriculture

Abstract. Soil moisture is crucial in various fields and monitoring it to guide irrigation represents a challenge. Machine learning has emerged as a promising tool to accurately predict soil moisture levels. This study focuses on evaluating machine learning techniques for this task, training models with meteorological variables and direct soil moisture measurements. Four machine learning algorithms were implemented, highlighting the Gradient Boosting Regressor as the most effective. In addition, a processed data set that combines meteorological and soil moisture measurements is presented, hoping that it will be useful for future research. This approach seeks to improve the compression and predictability of soil moisture, crucial for agricultural planning and water management in agriculture.

Keywords: Soil moisture, machine learning, regression models.

1. Introducción

El contenido de humedad del suelo es de vital importancia para una variedad de campos, incluyendo la biología, la hidrología, la agronomía, la ingeniería, la ecología y la geología del suelo. Su monitoreo es cada vez más extenso, especialmente con el incremento de inversiones en infraestructura de riego de precisión y sistemas de control. Sin embargo, la tarea de monitorear la humedad del suelo para guiar el riego presenta desafíos significativos. Los regantes deben seleccionar cuidadosamente el equipo adecuado para su sistema de riego y las características específicas de su parcela de tierra [1].

La humedad del suelo desempeña un papel crucial en el suministro de agua para la agricultura, siendo este su recurso principal. A pesar de su importancia, la medición directa en el campo enfrenta desafíos significativos, lo que subraya la necesidad de predecirla con precisión para respaldar actividades de planificación agrícola e investigaciones pertinentes [2].

El uso del aprendizaje automático ha dado lugar al desarrollo de algoritmos innovadores capaces de pronosticar de manera precisa los niveles de humedad del suelo, los cuales pueden ser empleados posteriormente en actividades de riego u otros propósitos [1]. Actualmente existen trabajos que aplican aprendizaje automático en la predicción de humedad del suelo. En [3] realizan la estimación de la humedad del suelo mediante aprendizaje profundo basado en datos satelitales.

Los autores en [4] utilizan la técnica de regresión denominada Máquina de Vectores de Soporte (SVM) para estimar la humedad del suelo mediante el uso de datos de teledetección. En el trabajo [5], se desarrollan y examinan modelos híbridos que combinan máquinas de aprendizaje extremo (ELM) con inteligencia de datos para realizar predicciones mensuales de humedad del suelo. En [6] hacen una estimación de la humedad del suelo basada en datos de teledetección y aprendizaje profundo.

Este estudio se enfoca en el uso de técnicas de aprendizaje automático para estimar la humedad del suelo. Motivado por encontrar una correlación entre datos meteorológicos y mediciones de humedad en el suelo dada la ausencia de sensores propios con datos abundantes que nos den una base sólida en cual sustentar otras investigaciones. Se entrenaron modelos con variables meteorológicas y mediciones directas de humedad del suelo.

Implementamos cuatro algoritmos de aprendizaje automático: Random Forest Regressor, K-Nearest Neighbors, Gradient Boosting Regressor y Regresión Lineal Múltiple. Al evaluar estos modelos con predicciones, encontramos que el Gradient Boosting Regressor demostró un error cuadrático medio y error absoluto medio menor en comparación con los otros modelos, especialmente al probar en intervalos de tiempo diferentes al de entrenamiento.

Finalmente, se aplicó este modelo a datos recientes de la estación meteorológica del Instituto Tecnológico de Morelia, obteniendo estimaciones de humedad del suelo consistentes y congruentes con el comportamiento esperado a lo largo del tiempo.

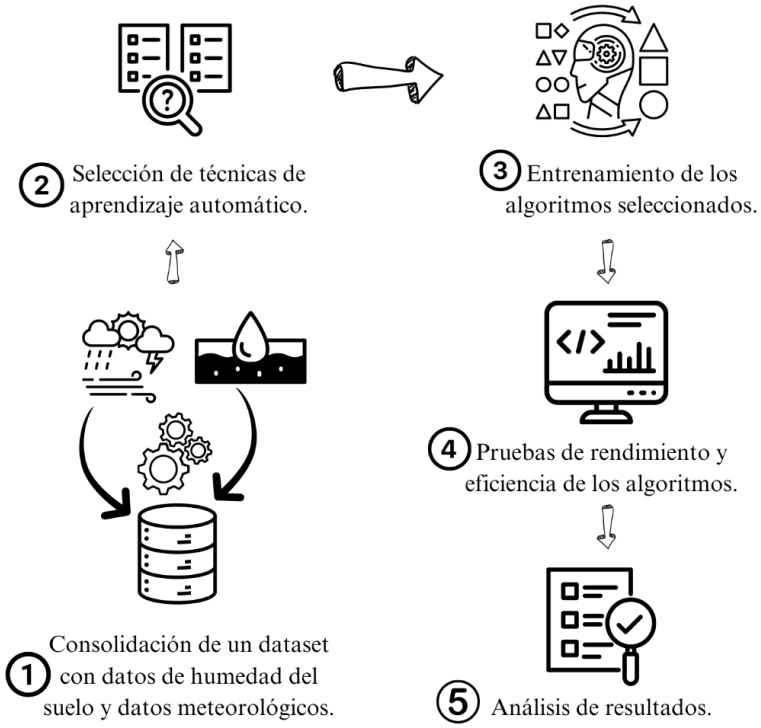

Fig. 1. Metodología.

Entre las contribuciones de este estudio, destacamos la presentación de un conjunto de datos creado por Gasch et al. [11]. Este conjunto de datos se procesó para ser presentado en un archivo CSV, extrayéndolo de su formato original en TXT.

Identificamos las ubicaciones de los sensores con la menor cantidad de datos faltantes y utilizamos el método de interpolación de vecinos más cercanos dado su estructura temporal para llenar los valores faltantes. Además, enriquecimos este conjunto de datos con información de una estación meteorológica cercana a las ubicaciones donde se tomaron las mediciones de humedad del suelo. De esta manera, creamos un conjunto de datos que integra mediciones meteorológicas y de humedad del suelo. Este conjunto de datos servirá de base para futuras investigaciones, ya sea para buscar patrones o realizar análisis temporales.

2. Marco teórico

2.1. Humedad del suelo

La agricultura y el agua están profundamente vinculadas, siendo el agua un factor esencial en la producción agrícola. Los métodos agrícolas influyen en el ciclo

Tabla 1. Variables encontradas en dataset meteorológico, definiciones y valores nulos.

Variable	Definición	Valores nulos
time	Tiempo	0
temp	Temperatura	213
dwpt	Punto de rocío	256
rhum	Humedad relativa	256
prcp	Precipitación	6558
snow	Profundidad de nieve	**80283**
wdir	Dirección del viento	23414
wspd	Velocidad media del viento	414
wpgt	Ráfaga máxima de viento	**80283**
pres	Presión	1125
tsun	Tiempo de sol	**80283**
coco	Código de condición climática.	**80283**
Total		**353368**

hidrológico mediante la evapotranspiración, la recarga de acuíferos y el flujo de aguas superficiales.

Una humedad adecuada en el suelo es crucial para varios procesos biológicos y físicos, incluyendo la germinación de semillas, el desarrollo vegetativo, el ciclo de nutrientes y la conservación de la biodiversidad del suelo. La medición de la humedad del suelo es esencial no solo para evaluar la disponibilidad de agua para la agricultura, sino también para entender la salud del suelo y su capacidad para retener agua, lo cual es vital para el mantenimiento de un agroecosistema sostenible [7].

La humedad del suelo es un factor crucial en la agricultura, pues influye directamente en el crecimiento de los cultivos y en la sostenibilidad de los ecosistemas agrícolas. Dicha humedad no solo depende de las prácticas de irrigación y del manejo del suelo, sino que está estrechamente vinculada a diversas variables climáticas.

2.2. Técnicas de aprendizaje automático

Las técnicas de aprendizaje automático, como Random Forest Regressor, K-Nearest Neighbors, Gradient Boosting Regressor y Regresión Lineal Múltiple son herramientas poderosas para predecir valores en una variedad de contextos. Estos algoritmos pueden emplearse para modelar relaciones complejas entre variables y generar predicciones precisas sobre valores futuros. Los autores en [8] presentan la aplicación de un método de aprendizaje automático en específico Random Forest Regressor para generar pronósticos diarios precisos de generación de energía solar, haciendo uso de datos históricos de mediciones y datos meteorológicos de fuentes abiertas suministrados por servicios meteorológicos.

En [9], se propone un método que utiliza el algoritmo de K-Vecinos Más Cercanos (KNN) para evaluar la calidad del suelo y predecir los cultivos más adecuados. Este enfoque considera la temperatura y la calidad del suelo como variables de entrada para el algoritmo. El artículo [10] describe el uso de modelos de aprendizaje automático para predecir la evapotranspiración de referencia, facilitando así la planificación del riego.

Fig. 2. Valor promedio mensual del dataset de variables meteorológicas.

Se emplearon datos meteorológicos diarios, que incluyen la temperatura máxima y mínima, humedad relativa, radiación solar, temperatura del suelo y velocidad del viento. Los datos se procesaron utilizando técnicas de Regresión Lineal Múltiple, Random Forest Regressor y Gradient Boosting Regressor Los resultados indicaron que el modelo preprocesado con GBR superó a los otros modelos en la precisión de las predicciones de evapotranspiración de referencia.

En el estudio [2], el objetivo principal fue predecir la humedad diaria del suelo a nivel de cultivo utilizando información meteorológica a través de modelos de regresión lineal múltiple. Se concluyó que estos modelos, al incorporar variables meteorológicas, resultaron efectivos para estimar la humedad del suelo. Esto se debe a que la humedad tendió a replicar los patrones estacionales y a responder a las variaciones en las precipitaciones. Que les falta a los demás trabajos o que tiene este trabajo que no tengan los mismos.

2.3. Datasets

En [11] presentan un conjunto de datos obtenidos del monitoreo del contenido de agua en el suelo, así como datos complementarios, recolectados en una granja experimental de labranza cero de 37 hectáreas situada en el noroeste de los Estados Unidos. Las mediciones del contenido de agua se han realizado cada hora desde el año 2007 mediante sensores ECH2O-TE y 5TE distribuidos en 42 ubicaciones, abarcando cinco profundidades (0.3; 0.6; 0.9; 1.2 y 1.5 metros), sumando un total de 210 sensores en toda la granja agronómica RJ Cook. Este conjunto de datos se encuentra disponible en [12].

Este conjunto de datos cuenta con mediciones horarias y diarias del contenido de agua (en m^3/m^3) y la temperatura del suelo (en °C) en 42 ubicaciones y en cinco profundidades (0.3; 0,6; 0,9; 1,2 y 1,5 metros) desde el 20 de abril de 2007 hasta el 16 de junio de 2016. Los datos se encuentran en archivos .txt para cada ubicación.

Tabla 1. Variables encontradas en dataset de humedad del suelo, definiciones y valores nulos.

Variable	Definición	Valores nulos
H_30cm	Humedad a 30 cm	18806
H_60cm	Humedad a 60 cm	23701
H_90cm	Humedad a 90 cm	22323
H_120cm	Humedad a 120 cm	24540
H_150cm	Humedad a 150 cm	25577
T_30cm	Temperatura a 30 cm	18806
T_60cm	Temperatura a 60 cm	23705
T_90cm	Temperatura a 90 cm	22330
T_120cm	Temperatura a 120 cm	24540
T_150cm	Temperatura a 150 cm	25578
Total		229906

El sitio web meteostat.net, es una base de datos meteorológicos y climáticos que proporciona datos detallados de miles de estaciones meteorológicas y lugares de todo el mundo. Afortunadamente, cuenta con una estación en Pullman, muy cerca de R.J. Cook Agronomy Farm, donde se tomaron las mediciones de contenido de agua del suelo y datos auxiliares a diferentes profundidades [13]. El sitio web [13], nos da la oportunidad de obtener datos de diferentes formas, más sin embargo cuando se descarga en un periodo de 7 días (una semana) los datos obtenidos tienen una frecuencia de cada hora lo cual es similar al dataset [12].

3. Metodología

La metodología empleada en este estudio se muestra en la Fig. 1. En primer lugar, se consolida un conjunto de datos que incluye humedad del suelo y datos meteorológicos (ver Sección 3.1). Luego, se seleccionan las técnicas de aprendizaje automático a utilizar. A continuación, se entrenan los algoritmos seleccionados para estimar la humedad del suelo. Posteriormente, se realizan pruebas y se evalúa la eficacia de los algoritmos elegidos. Finalmente, se analizan los resultados obtenidos.

3.1. Consolidar dataset que contenga la humedad del suelo y los datos meteorológicos

Esta etapa consta de consolidar un dataset que contenga la humedad del suelo del dataset [12] y los datos meteorológicos obtenidos de [13]. Primeramente, en el apartado 3.1.1 se muestra el proceso de obtención del dataset de datos meteorológicos.

3.1.1. Dataset de datos meteorológicos

El proceso de obtención de los datos meteorológicos consta de seleccionar el periodo de medición de 7 días a partir del día 20/4/2007 dentro de la estación de Pullman en el repositorio de meteostat.net, para después descargar el archivo seleccionando el

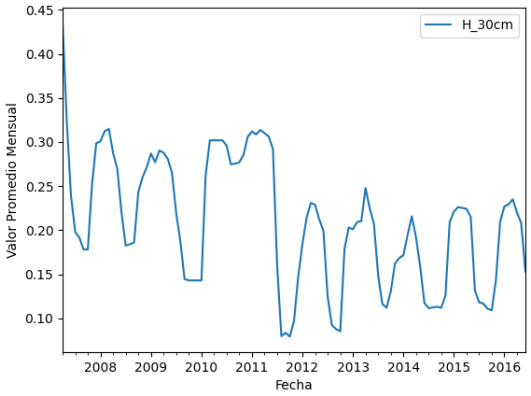

Fig. 1. Valores promedio mensuales de la humedad del suelo.

formato tipo CSV, de esta forma obtendremos archivos que contendrán la información de cada semana con frecuencia de cada hora. El proceso se repite hasta que llegamos a la fecha 16/6/2016, la cual compete al periodo del dataset [12] con el que se consolidara un nuevo dataset que contengan los datos de humedad del suelo y datos meteorológicos.

El primer reto fue la unión de todos los archivos descargados para la creación del dataset meteorológico, primero por año, y luego uno general que competerá a todo el periodo de prueba, para ello, generamos un código en Python que nos facilite la unión y ordenamiento de los datos de forma temporal.

En la Tabla 1 se muestra la información del dataset creado a partir de los datos meteorológicos, cuenta un número total de *80,283 registros* por variable y *12 variables*, se observa en la tabla que tanto las variables de snow, wpgt, tsun y coco, todos sus valores son nulos, por lo cual se eliminaran del dataset. Para las demás variables utilizaremos métodos para el llenado de estos valores nulos.

Existen varios métodos como: Imputación por media, mediana o moda, Regresión lineal o regresión múltiple, MICE (Multiple Imputation by Chained Equations), Matrix Factorization, Algoritmos de aprendizaje automático avanzados e Interpolación.

Dado que los datos están ordenados en el tiempo, es decir, tienen una estructura temporal, se puede utilizar métodos de interpolación para predecir los valores faltantes basados en los valores existentes.

El método de interpolación “nearest” o interpolación por vecino más cercano, es una forma de interpolación que se basa en la idea de que los valores cercanos en el tiempo (o en la secuencia) son más similares entre sí, por lo que el valor más cercano será una buena aproximación para el valor faltante. En la Fig. 2 se muestran las variables meteorológicas de este dataset.

3.1.2. Dataset de humedad del suelo

Como se mencionó anteriormente, este conjunto de datos incluye registros horarios y diarios de la humedad del suelo (expresada en m^3/m^3) y de la temperatura del suelo (en °C) en 42 ubicaciones diferentes y a cinco profundidades distintas (0.3; 0.6; 0.9; 1.2 y 1.5 metros). Estas mediciones se extienden desde el 20 de abril de 2007 hasta el 16 de junio de 2016. Los datos están almacenados en archivos de texto separados por ubicación. El primer desafío fue determinar cuál de estas ubicaciones ofrecía la mejor

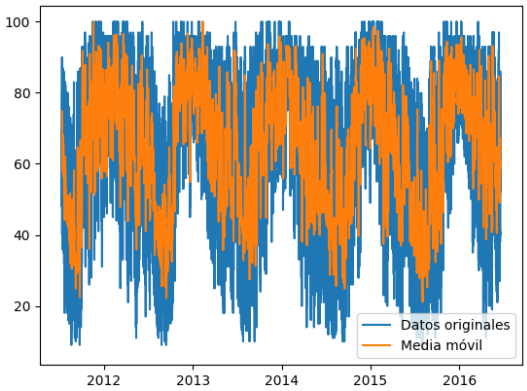

Fig. 2. Humedad relativa con y sin filtro.

calidad de datos, es decir, aquella que presentaba la menor cantidad de registros nulos. Esta selección fue crucial para asegurar la fiabilidad de los análisis subsiguientes.

Después de analizar los archivos de texto correspondientes, se determinó que el archivo CAF308.txt presentaba la menor cantidad de valores nulos en comparación con los archivos de las otras 42 ubicaciones de los sensores, específicamente en lo que respecta a las mediciones de humedad. En la Tabla 2 se detallan las variables del dataset de este archivo seleccionado para este estudio, como lo son la humedad del suelo y temperatura a diferentes profundidades, junto con la cantidad de valores nulos encontrados para cada variable.

Se observa que la medida de humedad del suelo a una profundidad de 30 cm es la que contiene menos valores nulos en comparación con las demás profundidades. El número total de valores de estas mediciones que se extienden desde el 20 de abril de 2007 hasta el 16 de junio de 2016 debería ser de *80283 registros*. Para abordar los valores faltantes, se empleó el método de interpolación "nearest" o interpolación por vecino más cercano. Al ser datos ordenados de forma temporal los valores cercanos en el tiempo tienden a ser más similares entre sí, lo que hace razonable suponer que el valor más próximo es una aproximación adecuada para el valor faltante.

En la Figura 3 se presenta el gráfico del valor promedio mensual de la humedad del suelo a una profundidad de 30 cm, que será el enfoque principal de este estudio. Se puede observar que abarca el período desde 2007 hasta 2016. Sin embargo, a simple vista se aprecia una tendencia más clara y significativa en los años comprendidos entre 2012 y 2016. Por lo tanto, para los análisis posteriores, nos centraremos en este intervalo de tiempo.

3.1.3. Dataset de humedad del suelo y datos meteorológicos

Después de adquirir los conjuntos de datos de humedad del suelo y datos meteorológicos, los combinamos en un único dataset. Luego, procedimos a crear una matriz de correlación entre las variables para explorar posibles relaciones. Dado que nuestra hipótesis sugería que la humedad relativa podría correlacionarse con la humedad del suelo, generamos un gráfico de la humedad relativa y aplicamos un filtro para suavizar el ruido. Utilizamos una media móvil con una ventana de tamaño 24, como se muestra en la Fig. 4.

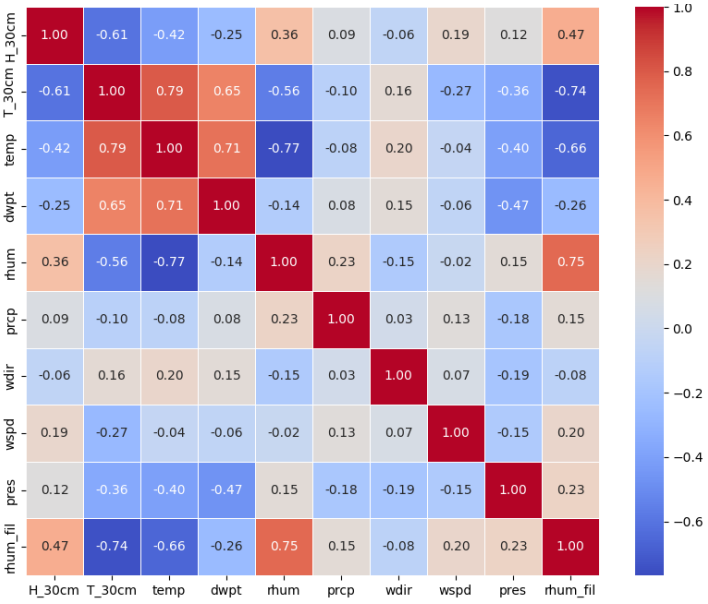

Fig. 5. Matriz de correlación de variables.

La ventana de tamaño 24 para la media móvil se seleccionó para coincidir con la periodicidad diaria de los datos recogidos cada hora, permitiendo cubrir un ciclo completo.

Este tamaño de ventana facilita el suavizado de las fluctuaciones diarias y resalta las tendencias claras en la humedad relativa, proporcionando una base sólida para analizar los efectos diarios en la humedad del suelo. En la Figura 3 se presenta la matriz de correlación entre las variables. Destaca que la correlación entre la Humedad del suelo (H_30cm) y la Humedad relativa (rhum) es de 0.36. Sin embargo, al aplicar el filtro de media móvil a la Humedad relativa, como se muestra en la Fig. 5, esta correlación aumenta a 0.47.

4. Resultados

Se realizó un análisis comparativo de cuatro técnicas de aprendizaje automático (Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors y Regresión Lineal Múltiple) para generar nuevos valores de humedad del suelo para fechas futuras, utilizando variables meteorológicas. Las variables seleccionadas fueron temperatura (temp), punto de rocío (dwpt), humedad relativa (rhum), precipitación (prcp) y el mes correspondiente. La selección de estas variables se justifica por la siguiente razón:

La humedad relativa mostró una correlación más alta con la humedad del suelo, como se observa en la Fig. 6. Además, dado el comportamiento cíclico observado en la Figura 6, se decidió incluir el mes como característica para el entrenamiento de los modelos. Las otras variables complementarias fueron seleccionadas porque están disponibles para futuros trabajos, utilizando datos meteorológicos de la ciudad de

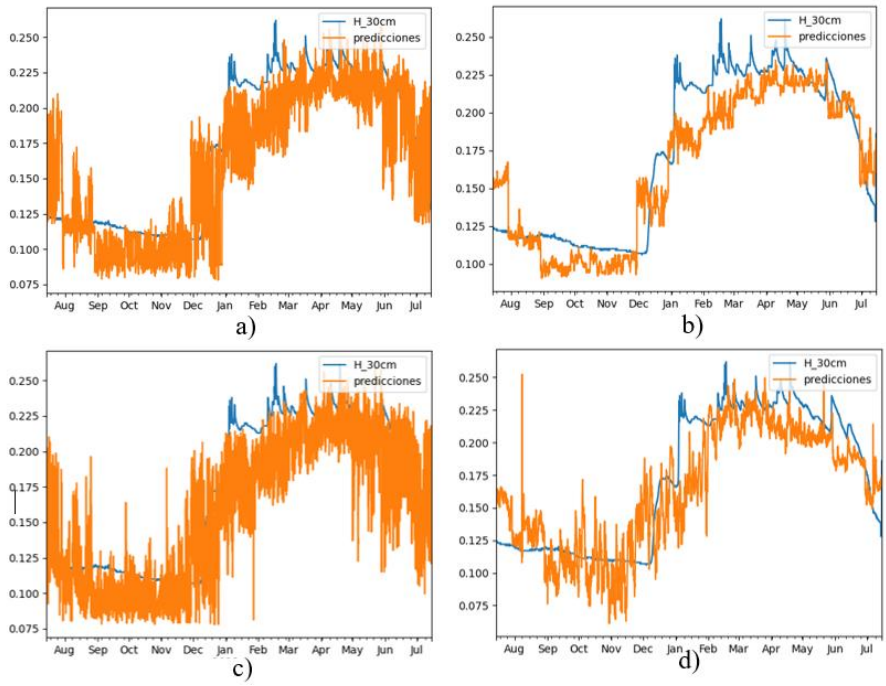

Fig. 3. Predicción de humedad del suelo de los diferentes modelos entrenados, a) Random Forest Regressor b) Gradient Boosting Regressor c) K-Nearest Neighbors y d) Regresión Lineal Múltiple.

Morelia, y serán utilizadas para generar valores sintéticos de humedad del suelo a partir del modelo entrenado

Para entrenar los modelos de aprendizaje automático, incluyendo Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors y Regresión Lineal Múltiple, se utilizó Python junto con la biblioteca Scikit-learn. Scikit-learn es una herramienta de código abierto que ofrece una amplia variedad de algoritmos tanto supervisados como no supervisados para el aprendizaje automático. En este estudio, se optó por emplear los valores predeterminados de los hiperparámetros para cada modelo, lo que garantiza una configuración estándar y coherente durante el proceso de entrenamiento.

Para entrenar el modelo, se empleó el conjunto de datos presentado en la sección 3.3, que incluye información sobre la humedad del suelo y datos meteorológicos. Para el proceso de entrenamiento, se decidió utilizar el período de tiempo comprendido entre 2011 y 2015, reservando el intervalo de 2015 a 2016 para probar el modelo.

Durante el entrenamiento con los datos de 2011 a 2015, se asignó el 80% de los datos para entrenamiento y el 20% restante para pruebas. Después de entrenar los modelos, se evaluaron utilizando datos del intervalo de 2015 a 2016 para probar sus predicciones. Se observó que el modelo Gradient Boosting Regressor mostró una mayor precisión al ajustarse a los valores reales, mientras que el modelo K-Nearest Neighbors exhibió mayores variaciones con respecto a los valores reales de la humedad del suelo.

Tabla 3. Error cuadrático Medio y Error Absoluto Medio de los modelos entrenados.

Variable	Entrenamiento ECM	Dataset 2015-2016 ECM	Entrenamiento EAM	Dataset 2015-2016 EAM
Random Forest Regressor	**0.000401**	0.000736	**0.013026**	0.020937
Gradient Boosting Regressor	0.000574	**0.000460**	0.018599	**0.017459**
K-Nearest Neighbors	0.000525	0.000814	0.015645	0.022255
Regresión lineal múltiple	0.001033	0.000718	0.025296	0.021316

Esta diferencia se puede verificar en la Tabla 3, donde se presentan las métricas de error cuadrático medio (ECM) y error cuadrático absoluto (ECA). Se observa un menor error al utilizar el modelo Gradient Boosting Regressor para estimar la humedad del suelo con datos distintos a los utilizados durante el entrenamiento.

Una vez que obtuvimos el modelo mejor evaluado, Random Forest Regressor, procedimos a realizar predicciones de la humedad del suelo utilizando los datos meteorológicos de la estación meteorológica del Instituto Tecnológico de Morelia. Para esto, necesitábamos la fecha de la cual se extrajo el mes, así como las variables de temperatura (temp), punto de rocío (dwpt), humedad relativa (rhum) y precipitación (prcp). En la Figura 7 se muestran los valores generados por nuestro modelo para el intervalo de tiempo de enero de 2021 a mayo de 2023. Se observa un comportamiento lógico y coherente, en línea con lo esperado.

5. Discusión de los resultados

Después de revisar exhaustivamente los resultados obtenidos, queda claro que el uso de técnicas de aprendizaje automático ofrece un rendimiento prometedor en el ámbito de la agricultura de precisión para estimar variables de interés. Al analizar los resultados de los modelos entrenados, se destaca que el Gradient Boosting Regressor demostró ser el más efectivo; sin embargo, las otras técnicas también arrojaron resultados positivos.

Para el entrenamiento de nuestros modelos, se emplearon los hiperparámetros preconfigurados por la biblioteca Scikit-learn, lo que proporcionó resultados satisfactorios. No obstante, se reconoce la posibilidad de mejorar la estimación utilizando técnicas para obtener hiperparámetros más específicos según el problema en cuestión.

En cuanto a las variables de entrada seleccionadas para el modelo, se optó por aquellas consideradas como la mejor opción dadas las limitaciones de enfoque y recursos del proyecto. No obstante, los resultados abren la puerta a la exploración de diferentes variables y enfoques, ya que se logró una estimación exitosa de la humedad del suelo utilizando técnicas de aprendizaje automático. Dado que no contamos con sensores de humedad en todas las regiones de estudio, debido a los costos asociados, es valioso contar con técnicas que proporcionen estimaciones precisas de variables específicas.

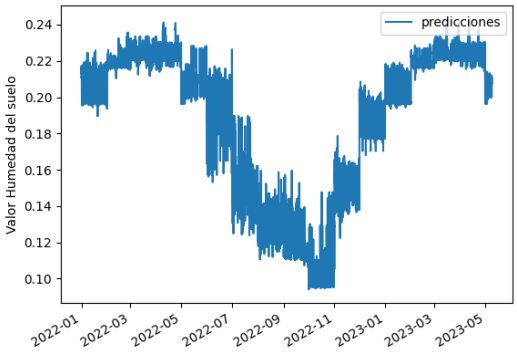

Fig. 4. Predicción de la humedad del suelo para dataset meteorológico del ITM.

Por esta razón, al tener solo datos meteorológicos y no mediciones directas de humedad del suelo, fue de interés realizar estimaciones basadas en modelos generados con aprendizaje automático a partir del comportamiento de dichas variables en diferentes ubicaciones. Este enfoque abre la oportunidad de experimentar con más variables o técnicas, así como de adquirir y tratar datos adicionales de interés en la agricultura. Además, una vez que se disponga de datos reales de medición de humedad del suelo en el área de Morelia, será posible compararlos con las estimaciones realizadas por el modelo implementado aquí.

6. Conclusiones

Este trabajo demuestra la aplicabilidad de las técnicas de aprendizaje automático para la estimación de la humedad del suelo, con potencial para futuras investigaciones y aplicaciones en el campo agrícola. Se exploraron y compararon diversas técnicas de aprendizaje automático para la estimación de la humedad del suelo en el contexto de la agricultura. A través del análisis de los resultados obtenidos, se pudo observar el potencial de estas técnicas para proporcionar estimaciones precisas y útiles en la planificación agrícola y la toma de decisiones.

Los modelos entrenados utilizando Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors y Regresión Lineal Múltiple demostraron su eficacia para predecir la humedad del suelo utilizando variables meteorológicas como temperatura, punto de rocío, humedad relativa, precipitación y el mes correspondiente. Entre estos modelos, el Gradient Boosting Regressor se destacó por su menor error cuadrático medio y error absoluto medio, lo que sugiere su mayor capacidad predictiva en comparación con las otras técnicas evaluadas. Además, se observó que la inclusión del mes como característica en el entrenamiento de los modelos contribuyó significativamente a mejorar su rendimiento, lo que indica la importancia de considerar la variabilidad estacional en la estimación de la humedad del suelo.

El procesamiento y análisis de datos realizado en este estudio también proporcionó perspectivas importantes sobre la disponibilidad de información y la viabilidad de las técnicas de aprendizaje automático en entornos agrícolas donde los datos de sensores pueden ser limitados o costosos de adquirir.

Otra contribución este trabajo es la presentación y procesamiento del conjunto de datos creado por C. K. Gasch et al. Este conjunto de datos se ha transformado en un formato más accesible, facilitando su uso y análisis para futuras investigaciones en el campo de la estimación de la humedad del suelo y la agricultura de precisión. Al identificar y abordar los datos faltantes utilizando técnicas de interpolación de vecinos más cercanos y enriquecer el conjunto de datos con información meteorológica adicional, hemos creado una base sólida para análisis más detallados y comprensivos.

Agradecimientos. Los autores agradecen al Tecnológico Nacional de México por el apoyo brindado a través del proyecto 19476.24-P. Noel A. Zavala-Díaz agradece al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de posgrado 2021-000018-02NACF-12060 brindada a través del Instituto Tecnológico de Morelia.

Referencias

1. Rasheed, M.W., Tang, J., Sarwar, A., Shah, S., Saddique, N., Khan, M.U., Imran-Khan, M., Nawaz, S., Shamshiri, R.R., Aziz, M., Sultan, M.: Soil Moisture Measuring Techniques and Factors Affecting the Moisture Dynamics: A Comprehensive Review. Sustainability, vol. 14, no. 18, pp. 11538 (2022). DOI: 10.3390/su141811538.
2. Palominos-Rizzo, T., Villatoro-Sánchez, M., Alvarado-Hernández, A., Cortés-Granados, V., Paguada-Pérez, D.: Estimación de la humedad del suelo mediante regresiones lineales múltiples en llano brenes, Costa Rica. Agronomía Mesoamericana, pp. 47872 (2022). DOI: 10.15517/am.v33i2.47872.
3. Lee, C., Sohn, E., Park, J., Dong, J.J.: Estimation of Soil Moisture using Deep Learning based on Satellite Data: A Case Study of South Korea. GIScience & Remote Sensing, vol. 56, no. 1, pp. 43–67 (2018). DOI: 10.1080/15481603.2018.1489943.
4. Ahmad, S., Kalra, A., Stephen, H.: Estimating Soil Moisture using Remote Sensing Data: A Machine Learning Approach. Advances in Water Resources, vol. 33, no. 1, pp. 69–80 (2010). DOI: 10.1016/j.advwatres.2009.10.008.
5. Prasad, R., Deo, C., Li, Y., Maraseni, T.: Soil Moisture Forecasting by a Hybrid Machine Learning Technique: ELM Integrated with Ensemble Empirical Mode Decomposition. Geoderma, vol. 330, pp. 136–161 (2018). DOI: 10.1016/j.geoderma.2018.05.035.
6. Wang, G., Hu, P., Lai, X., Xue, B., Fang, Q.: Root-Zone Soil Moisture Estimation based on Remote Sensing Data and Deep Learning. Environmental Research, vol. 212, pp. 113278 (2022). DOI: 10.1016/j.envres.2022.113278.
7. Kashyap, B., Kumar, R.: Sensing Methodologies in Agriculture for Soil Moisture and Nutrient Monitoring. IEEE Access, vol. 9, pp. 14095–14121 (2021). DOI: 10.1109/access.2021.3052478.
8. Khalyasmaa, A., Eroshenko, S.A., Chakravarthy, T.P., Gasi, V.G., Bollu, S.K.Y., Caire, R., Atluri, S.K.R., Karrolla, S.: Prediction of Solar Power Generation based on Random Forest Regressor Model. In: International Multi-Conference on Engineering, Computer and Information Sciences, pp. 0780–0785 (2019). DOI: 10.1109/sibircon48586.2019.8958063.
9. Gajula, A.K., Singamsetty, J., Dodda, V.C., Kuruguntla, L.: Prediction of Crop and Yield in Agriculture using Machine Learning Technique. In: 12th International Conference on Computing Communication and Networking Technologies, pp. 1–5 (2021). DOI: 10.1109/icccnt51525.2021.9579843.
10. Ponraj, A.S., Vigneswaran, T.: Daily Evapotranspiration Prediction using Gradient Boost Regression Model for Irrigation Planning. The Journal of Supercomputing, vol. 76, no. 8, pp. 5732–5744 (2019). DOI: 10.1007/s11227-019-02965-9.

11. Gasch, C.K., Brown, D.J., Campbell, C.S., Cobos, D.R., Brooks, E.S., Chahal, M., Poggio, M.: A Field-Scale Sensor Network Data Set for Monitoring and Modeling the Spatial and Temporal Variation of Soil Water Content in a Dryland Agricultural Field. Water Resources Research, vol. 53, no. 12, pp. 10878–10887 (2017). DOI: 10.1002/2017wr021307.
12. Gasch, C.K., Brown, D.J., Campbell, C.S., Cobos, D.R., Brooks, E.S., Chahal, M., Poggio, M.: A Field-Scale Sensor Network Data Set for Monitoring and Modeling the Spatial and Temporal Variation of Soil Water Content in a Dryland Agricultural Field. Water Resources Research, vol. 53, no. 12, pp. 10878–10887 (2017). DOI: 10.1002/2017wr021307.
13. Meteostat: Pullman/Sunshine. https://meteostat.net/es/station/KPUW0?t=2007-04-20/2007-04-27 (2024)