UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Maestría en Ciencia de Datos y Analítica

Proyecto integrador 1

2024-1

Análisis de la humedad del suelo usando datos meteorológicos

Grupo 10

Autores:

Manuela Ramos Ospina[1]

Camila Acosta Gómez[2]

John Zapata Jimenez[3]

Dany Palacio Agudelo[4]

[1] Universidad EAFIT, Escuela de Ciencias Aplicadas e Ingeniería. Pasante de investigación. mramoso@eafit.edu.co

[2] Yamaha. Científica de datos. acacostag@eafit.edu.co

[3] Bancolombia. Científico de datos. jfzapataj@eafit.edu.co

[4] Servicios Ambientales y Geográficos S.A. Medio biótico. Asistente de coordinación. dcpalacioa@eafit.edu.co

Tabla de Contenido

1. Introducción

La humedad del suelo es una variable que determina el estado hidrológico del suelo y a su vez relaciona lo que sucede en la superficie de la tierra con procesos atmosféricos. El conocimiento preciso de esta variable y sus dinámicas temporales son de crucial importancia para gran cantidad de aplicaciones, como la predicción climática y meteorológica, la gestión agrícola, la distribución de los recursos hídricos, la prevención de desastres naturales y numerosos problemas geotécnicos[5].

El estado de la humedad del suelo en sus capas más superficiales se caracteriza por una alta movilidad espacial y temporal, influenciada por variables meteorológicas como la precipitación, temperatura, radiación solar, entre otros[6]. Pero también depende de variables geofísicas como las propiedades del suelo, características topográficas y de vegetación. Es por esto que su estimación y monitoreo a larga escala es un reto[7].

a. Problema de negocio

En general, la estimación de la humedad del suelo se realiza mediante tres enfoques: (1) observaciones in situ, (2) sensado remoto y (3) modelos empíricos. Las observaciones in situ proveen información localizada, en escalas de tiempo que van desde los minutos hasta largos períodos, usando sensores o muestreo destructivo. Por otro lado, el sensado remoto realizado con satélites, tiene escalas de tiempo que van desde días hasta semanas. Sin embargo, en muchas aplicaciones prácticas no hay sensores disponibles o los datos satelitales no proveen la resolución espacial o temporal necesaria[5]. Es por esto, que los modelos empíricos se han posicionado como una herramienta alternativa para estimar la humedad del suelo e incluso interpretar las relaciones entre uno o más predictores a partir de datos que se puedan obtener fácilmente.

b. Objetivos

[5] ANELLO, Mirko, BITTELLI, M., BORDONI, M. Robust Statistical Processing of Long-Time Data Series to Estimate Soil Water Content. Mathematical Geosciences, 2024, vol. 56, no 1, p. 3-26.

[6] PALOMINOS-RIZZO, Teresa, VILLATORO-SÁNCHEZ, Mario, ALVARADO-HERNÁNDEZ, Alfredo, CORTÉS-GRANADOS, Víctor, & PAGUADA-PÉREZ, Darwin. Estimación de la humedad del suelo mediante regresiones lineales múltiples en Llano Brenes, Costa Rica. Agronomía Mesoamericana, 2022, vol. 33, no 2, p. 13.

[7] BABAEIAN, Ebrahim, M. SADEGHI, S. B. JONES, C. MONTZKA, H. VEREECKEN, and M. TULLER. Ground, proximal, and satellite remote sensing of soil moisture. Reviews of Geophysics, 2019, vol. 57, no 2, p. 530-616., doi: 10.1029/2018RG000618.

Aunque hay un interés generalizado en expandir el horizonte de comprensión de la humedad del suelo, poco se ha discutido acerca del monitoreo de esta variable usando datos meteorológicos, teniendo en cuenta que las redes de sensores meteorológicas proveen información constante y en muchos casos de forma pública. Es por esto que en este proyecto se plantea analizar la humedad del suelo y su relación con variables meteorológicas a partir de la recolección, curado, y transformación de una base de datos y su consecuente modelado estadístico. Además de la generación de un tablero en el qué visualizar los resultados para su uso y aplicación por parte del usuario.

c. Impacto de la solución

El conocimiento de la humedad del suelo tiene impacto y usos significativos en diversas áreas. En particular, desde la línea de investigación en Instrumentación Electrónica de la Universidad EAFIT, actualmente se busca implementar instrumentos para la medición de la humedad del suelo. Sin embargo, se observa la necesidad de generar modelos para estimarla, dado el alto costo de los sensores comerciales, reto en la fabricación masiva de sensores de bajo costo y la complejidad de calibración de los sensores existentes. En contraste, las variables meteorológicas han sido ampliamente estudiadas y su medición es habitual. Basta con mencionar que el Sistema de Alerta Temprana del Valle de Aburrá (SIATA) reporta una disposición de 35 sensores meteorológicos activos en el Valle de Aburrá, en contraste con los 9 sensores de humedad del suelo.

Es por esto que el análisis de la humedad del suelo de forma indirecta con las variables ambientales sirve como herramienta de apoyo para tener más conocimiento de esta variable, de forma remota y en sitios donde no se tiene facilidad de acceso para la medición directa.

2. Marco teórico

a. Humedad del suelo

La humedad del suelo (HS) representa la cantidad de agua presente en un suelo y se relaciona con las fuerzas que mantienen el agua dentro del sólido. La HS se expresa generalmente como un porcentaje que representa la cantidad de agua en relación con el peso o volumen total del suelo, llamado contenido de humedad del suelo[7].

Es posible medir la HS de forma directa o indirecta. Los métodos directos implican la extracción de una muestra del suelo para su análisis con procesos estándar. Aunque son precisos, estos métodos son destructivos y laboriosos. Por otro lado, los métodos indirectos evalúan la interacción del contenido de agua en el suelo con otros fenómenos, como su respuesta ante ondas electromagnéticas, cambios de presión y lo que nos interesa en este trabajo: eventos meteorológicos.

Si bien la HS está ligada a múltiples tipos de variables tanto naturales como artificiales, la parte poco profunda del suelo es la más afectada por las variables atmosféricas, siendo la precipitación el factor más estudiado en este balance hídrico, ya que determina la cantidad de agua que ingresa al

suelo[8]. Otras variables que influyen en la evolución del suelo son la temperatura (relacionada a su vez con la radiación solar), el viento y la humedad relativa, en la medida en que éstas condicionan los procesos de evapotranspiración[9].

b. Estándar para el procesamiento analítico de los datos

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) es un marco de trabajo estándar utilizado en minería de datos que proporciona una estructura para el desarrollo de modelos analíticos. Durante dos décadas, la metodología CRISP-DM ha influenciado otros estándares o a servido como punto de partida para la aplicación selectiva de esta metodología en otros procesos[10].

La metodología CRISP-DM se conceptualiza en 6 fases, las cuales fueron adaptadas al contexto específico de este proyecto y se esquematizan en la Figura 1.

[8] FARAGO, T. Soil moisture content: Statistical estimation of its probability distribution. Journal of Applied Meteorology and Climatology, 1985, vol. 24, no 4, p. 371-376.

[9] JARAMILLO, Daniel F. Introducción a la ciencia del suelo. 2002.

[10] HAYA, Pablo. La metodología CRISP-DM en ciencia de datos. Instituto de Ingeniería de Conocimiento. 2021. Accedido: may. 12, 2024 [En línea]. Disponible en: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

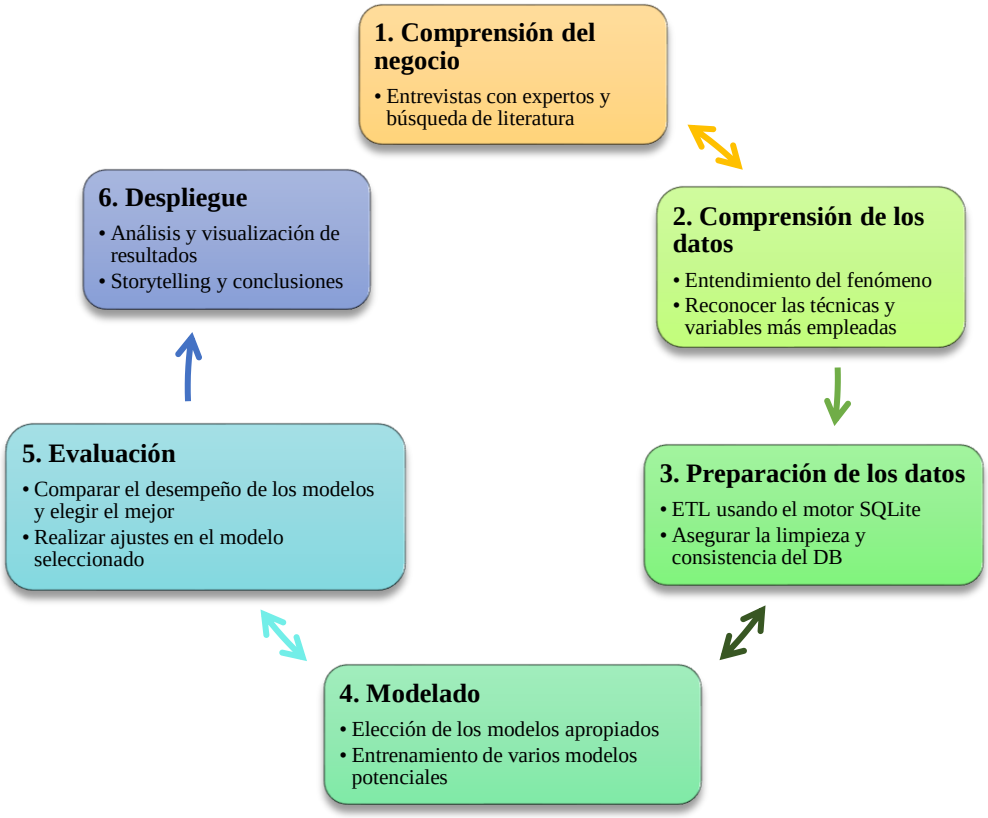


Figura 1. Metodología para el desarrollo del proyecto basada en CRISP-DM

c. Estado del arte

Es necesario examinar la literatura existente para entender los avances, desafíos y tendencias en el problema de investigación. A partir de la recopilación de literatura en motores de búsqueda académicos y su consecuente curado, se seleccionaron seis artículos en los que se observa que los principales métodos para estudiar la humedad del suelo usando datos meteorológicos se basan en algoritmos de aprendizaje estadístico y redes neuronales. Los modelos usados incluyen Redes Neuronales Simples[11], Redes de Memoria a Largo Plazo[12], algoritmos de aprendizaje de máquina

[11] PRAKASH, Shikha; SAHU, Sitanshu Sekhar. Soil moisture prediction using shallow neural network. International Journal of Advanced Research in Engineering and Technology, 2020, vol. 11, no 6. Available at SSRN: https://ssrn.com/abstract=3656915

[12] ORTH, Rene, et al. High-resolution European daily soil moisture derived with machine learning (2003–2020). Scientific Data, 2022, vol. 9, no 1, p. 1-13. doi: 10.1038/s41597-022-01785-6. PMID: 36376361; PMCID: PMC9663700

como Máquinas de Soporte Vectorial, Bosques Aleatorios[13] y Máquinas de aprendizaje extremo[14]. Mientras que solo dos artículos se enfocan en métodos estadísticos como Regresiones Lineales[6] y estadísticas robustas paramétricas y no paramétricas[5].

Además, se reconoció que las variables con mayor frecuencia de entrada para el análisis de la humedad del suelo fueron: temperatura, precipitación, humedad relativa y velocidad del viento. Lo anterior es especificado en la Tabla 1.

Tabla 1. Comparación de los métodos y variables usadas en la literatura consultada

Referencia	Titulo	Año	Métodos	Variables de entrada
(Hussain et al., 2023)	Estimation of Soil Moisture with Meteorological Variables in Supervised Machine Learning Models	2023	support vector regression (SVR) and random forest (RF)	Temperatura, humedad relativa, velocidad del viento, y precipitación
(Prakash & Sahu, 2020)	Soil Moisture Prediction Using Shallow Neural Network	2020	multiple linear regression, support vector regression and shallow neural network	Sin datos disponibles
(Feng et al., 2019).	Estimation of soil temperature from meteorological data using different machine learning models	2019	extreme learning machine (ELM), generalized regression neural networks (GRNN), backpropagation neural networks (BPNN) and random forests (RF)	Temperatura ambiente, velocidad del viento, humedad relativa, radiación solar y déficit de la presión de vapor
(Anello et al., 2024)[15]	Robust Statistical Processing of Long-Time Data Series to Estimate Soil Water Content	2024	robust statistics: parametric and non-parametric models	Precipitación y temperatura ambiente

[13] HUSSAIN, Mariam; SHARMIN, Nusrat; SHAFIUL, Sumayea Binte. Estimation of Soil Moisture with Meteorological Variables in Supervised Machine Learning Models. En 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2023. p. 1-6. doi: 10.1109/ECCE57851.2023.10101650

[14] FENG, Yu, et al. Estimation of soil temperature from meteorological data using different machine learning models. Geoderma, 2019, vol. 338, p. 67-77. https://doi.org/10.1016/J.GEODERMA.2018.11.044

[15] ANELLO, Mirko, et al. Robust Statistical Processing of Long-Time Data Series to Estimate Soil Water Content. Mathematical Geosciences, 2024, vol. 56, no 1, p. 3-26.

(O et al., 2022)	High-resolution European daily soil moisture derived with machine learning (2003-2020)	2022	long short-term memory networks	Temperatura ambiente, precipitación radiación superficial neta, temperatura superficial, acidez, topografía, cobertura terrestre, tipo de suelo
(Palominos-Rizzo et al., 2022)[16]	Estimación de la humedad del suelo mediante regresiones lineales múltiples en Llano Brenes, Costa Rica	2022	Multiple linear regression, PCA + linear regression	Temperatura ambiente, humedad relativa, velocidad del viento y precipitación acumulada

3. Desarrollo metodológico

a. Recuperación de los datos

Disponibilidad de la información

Dados los acercamientos con los diferentes expertos respecto a disponibilidad de la información y métodos de medición de las variables de interés, se recibieron sugerencias respecto a la recuperación del histórico requerido, a los tipos de análisis que podrían ser viables y a las posibles fuentes de información a usar. Luego de un proceso de selección se elige las fuentes de datos del SIATA dado a su acceso público, cubrimiento a nivel del área metropolitana y disponibilidad histórico1 de los datos.

Al realizar la búsqueda de los datos en el SIATA, se encontró que las estaciones en las que se mide la variable humedad del suelo, no se miden las demás variables meteorológicas y que además, hay algunos meses para los que se presentan datos faltantes, por lo que se tuvo un acercamiento con la experta Juliana Álvarez, quien confirmó la estación de la cual podríamos obtener información de todas las variables de interés, y quien además mencionó que actualmente, los puntos con datos faltantes son un tema en el que en la entidad se está trabajando para resolver, por lo que los resultados que se puedan tener en ese sentido podrían ser de utilidad para el proyecto SIATA.

Dado que la principal variable de interés es la humedad del suelo, pero que no está disponible para su descarga desde el portal, se realizó una solicitud formal a la entidad, desde donde se obtuvo que las estaciones de la siguiente tabla son las que cuentan con la medición para dicha variable.

[16] PALOMINOS-RIZZO, Teresa, et al. Estimación de la humedad del suelo mediante regresiones lineales múltiples en Llano Brenes, Costa Rica. Agronomía Mesoamericana, 2022, vol. 33, no 2, p. 13.

Tabla 2. Estaciones con Medición de humedad del suelo

Código	Estación	Latitud	Longitud	Ciudad	Vereda	Fecha de instalación
296	Laguna Quitasol - Humedad	6.35787	-75.555	Bello	Tierradentro	23/04/2018
396	Quebrada Larga Stevens	6.20556	-75.685	Medellín	Yarumalito	3/10/2019
532	Finca Altamira - Humedad	6.13848	-75.622	Sabaneta	Cañaveralejo	19/05/2021
541	Vereda Potrerito - Humedad	6.42564	-75.32	Barbosa	Potrerito	18/06/2021
553	La Sanin - Humedad Stevens	6.20326	-75.546	Medellín	Las Palmas	14/09/2021
608	I. E. San Luis Gonzaga Sede Carlos Mesa - Humedad Stevens	6.31511	-75.494	Copacabana	Las Margaritas	8/06/2022
611	Laguna Quitasol - Humedad Stevens	6.35789	-75.555	Bello	Tierradentro	14/06/2022
613	Vereda Naranjal - Humedad Stevens	6.30212	-75.659	Medellín	Naranjal	17/06/2022
620	Cabana Musical - Humedad Stevens	6.205	-75.487	Medellín	Santa Elena - Sector Central	29/06/2022
680	La Valeria - Humedad Stevens	6.09136	-75.643	Caldas	La Valeria	14/09/2023

Con base a la tabla anterior se contrastó la información disponible de esta variable, frente a las demás de interés. Se encontró que, solo para 4 de estas hay otra estación cercana en la que se miden variables meteorológicas así:

- Estación de humedad del suelo 532-Finca Altamira cerca a la estación meteorológica 318-Rafael Mejía. Al analizar los datos de humedad del suelo en esa estación, se encontró la mayoría de datos faltantes, por lo que se tuvo que descartar.
- Estación de humedad del suelo 296-Quitasol, cercana a la estación meteorológica 271-Jorge Eliecer Gaitán. Esta estación tiene 20% de datos nulos entre las fechas 8-07-2022 al 30-04-2024, y además la estación meteorológica asociada se encuentra en el centro de la ciudad de Bello por lo que añadiría variables adicionales no contempladas, por tanto se considera inadecuada.

- Estación de humedad 396-Quebrada Larga: Tiene solo 13% de datos nulos para las fechas 8-07-2022 al 30-04-2024 pero se encuentra muy lejos (~ 6.23 km) de la estación meteorológica más cercana.
- **Estación de humedad 541-Vereda Potrerito**: Cercana (~ 1.20 km) a la estación meteorológica 82-IE Manuel Jose Caicedo. Dada su ubicación en las afueras de la ciudad se considera la más adecuada para el análisis.

Lo anterior se puede ver en la siguiente imagen.

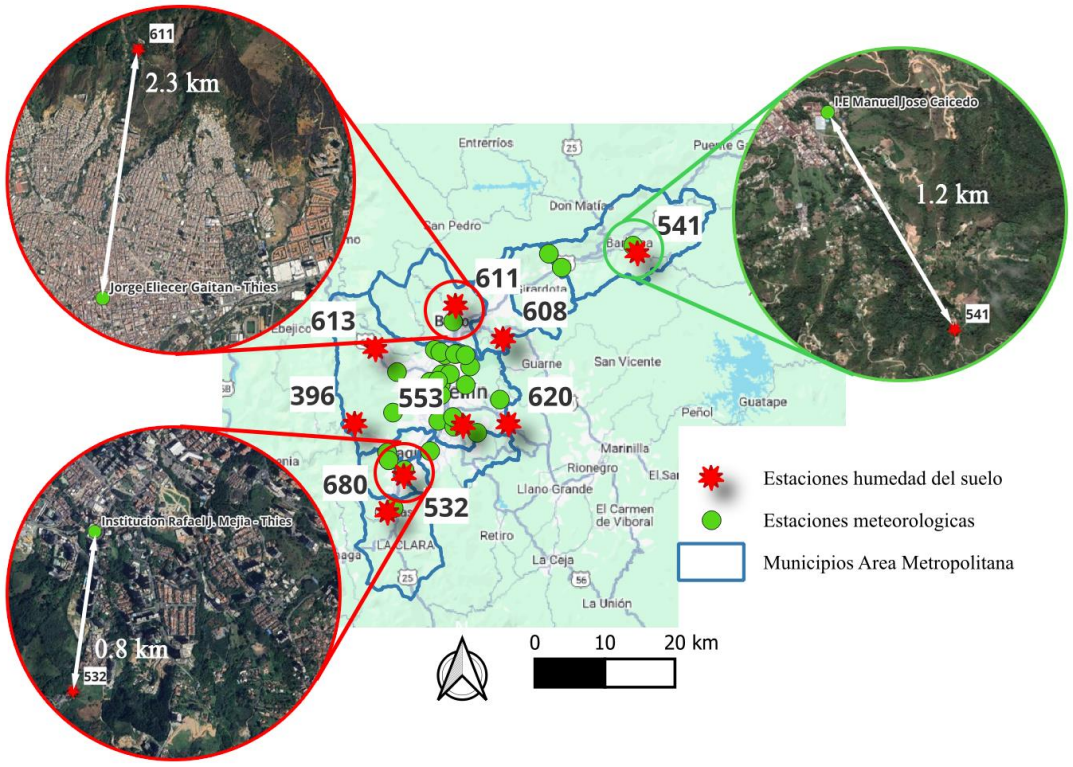


Figura 2. Ubicación y distancia entre las estaciones meteorológicas y las estaciones de medición de la humedad del suelo [Imagen de construcción propia].

Respecto a la variable humedad del suelo es importante mencionar que en la información brindada se tienen datos de diferentes alturas de medición llamados $h_1 = 0.05m$, $h_2 = 0.5m$ y $h_3 = 1m$. Al revisar los datos, se encontró que solamente se tenían suficiente información de altura h_2 para hacer viable el análisis. Idealmente se debería tomar h_1, la altura más superficial, ya que tiene mayor dependencia con las variables atmosféricas, pero en este caso, la elección de profundidad se eligió por la información disponible.

Una vez seleccionada la estación de interés y con la información de la humedad disponible por el SIATA, se procedió a crear la sabana de datos.

ETL

Se desarrolló un script de ingesta en batch automática que recibe la carpeta donde se ingestan los archivos y el nombre deseado para la creación de la respectiva tabla y su almacenamiento en una base de datos. Este fue ejecutado usando el método de conexión a la base de datos SQLite.

Ingeniería de datos

Una vez construidas las tablas, se posibilita la exploración de los datos con el lenguaje de programación SQL, esto dado que las tablas cuentan datos estructurados.

Así entonces, se construyó un script de SQL para encontrar la fecha mínima y máxima de cada variable de la base de datos y de esta forma crear una nueva tabla con dicha dimensión temporal a una frecuencia de 1 hora, llamada "fecha_hora". Ya que las variables originalmente tienen una escala temporal de minutos, esto se logra promediando los 60 valores.

Adicionalmente se construyen variables exógenas a partir de la fecha de cada registro: "mes", "día_mes", "semana" y "día". Finalmente, esta tabla es adjuntada a la base de datos.

Preparación de los datos

Una vez completada la ingesta de la dimensión de tiempo a la base de datos local, se continúa con la construcción del modelo multidimensional. Se realizó una limpieza inicial de los datos eliminando los registros cuyo índice de calidad fuera dudosa para la respectiva variable. De esta manera se dejaron únicamente los índices de calidad 1 y 2 asociados con "Calidad confiable del dato en tiempo real" y "Calidad confiable del dato no obtenido en tiempo real", respectivamente, así como fue especificado por el SIATA.

Despliegue y creación de la base de datos

Finalmente, se desarrolló un Query integrador que carga la base de datos completa para el análisis y construcción del modelo. A continuación se puede observar el resultante de la base de datos después de completar el proceso de ETL.

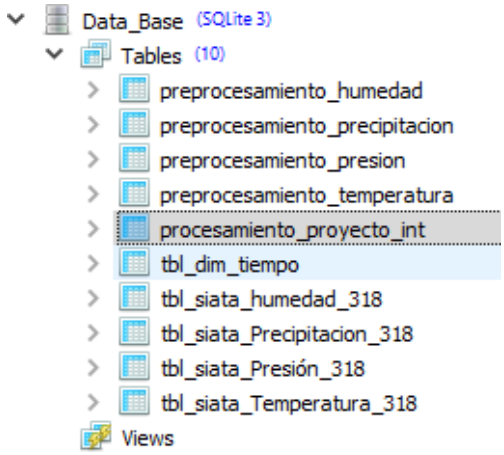


Figura 3. Base de datos resultante en SQLite 3

b. Análisis exploratorio

Descripción de los datos

El conjunto de datos producto del proceso ETL consta de diez (10) variables ambientales: nueve predictoras y la variable respuesta. Estas son especificadas en la Tabla 3Tabla 3.

Tabla 3 Variables ambientales contempladas en el conjunto de datos producto del proceso ETL

Índice	Nombre de la variable	Abreviatura	Unidades
1	Precipitación	Prec	mm
2	Presión atmosférica	P	hPa
3	Temperatura	T	°C
4	Humedad relativa / Humedad	H	%
5	Magnitud de la velocidad promedio del viento	V. Prom.	m/s
6	Magnitud de la Velocidad Máxima del viento	V. Max.	m/s
7	Dirección promedio del viento	Dir. Prom.	grados
8	Dirección Máxima del viento	Dir. Max.	grados
9	Radiación solar	Rad.	W/m^2
10	Contenido de humedad del suelo / Humedad del suelo	CH	$m^3/\ m^3$

Además de estas diez (10) también se incluyeron las cinco (5) variables relacionadas con la temporalidad que fueron introducidas en la sección Ingeniería de datos.

En particular, en la Figura 4 se representa la distribución y gráfico de cajas y bigotes (boxplot) para cada variable de la Tabla 3. Se puede observar que la precipitación es una variable altamente sesgada por naturaleza, puesto que lo más común es que en un día (24 horas que equivalen a 24 registros) los eventos de lluvia sean reportados en unas pocas horas, por tanto, la mayor concentración de los datos será en 0.0 mm en promedio por hora. Mientras que cuando llueve la intensidad es muy variable, lo que conduce a una gran dispersión de los datos en el resto de valores, cuyo máximo es de ~0.8 mm en promedio por hora. De forma similar se comporta la radiación solar[17] (la cual está asociada directamente con la constante solar, que es el flujo de energía proveniente del Sol que llega a la atmósfera terrestre de manera perpendicular y es de ~ 1,367 W/m^2, la cual constituye una asíntota para esta variable), siendo altamente sesgada con una alta concentración de valores en cero (en las noches la radiación solar es nula) y alta dispersión en el resto de los valores, llegándose a elevar hasta los 1,178 W/m^2, sin alcanzar nunca la magnitud de la constante solar. Además, en cuanto al viento, al ser un vector, se mide tanto su magnitud como su dirección máxima y promedio. Las velocidades máximas y promedios del viento son claramente sesgadas a la derecha y unimodales. Mientras que las direcciones promedio y máximas son bimodales y también sesgadas a la derecha, lo cual indica que eventualmente podrían ser variables redundantes.

[17] EL MGHOUCHI, Youness. Solar energy modelling and forecasting using artificial neural networks: a review, a case study, and applications. En Artificial Neural Networks for Renewable Energy Systems and Real-World Applications. Academic Press, 2022. p. 113-147. 2

Por su parte, la temperatura registra valores desde los 14 °C hasta los 30 °C, teniendo mayor concentración de datos en torno a los 18 °C. Mientras que la humedad relativa siendo un porcentaje, abarca casi totalmente una centena, sin tocar los valores extremos, pero mayoritariamente sesgada a valores altos, siendo esto común en el trópico. En oposición a las otras distribuciones, la presión es la que expresa un comportamiento aproximadamente normal, con valores que rondan los 864 hPa, y no presenta mayor variabilidad pues ésta depende principalmente de la altura, y en menor medida de la temperatura y la densidad del aire.

Finalmente, la variable respuesta es adimensional y toma valores que entendidos como porcentajes, donde 0% corresponde al espacio vacío (nada de agua en el suelo) y 100% correspondería a agua. En particular se registran valores entre 20 y 50.

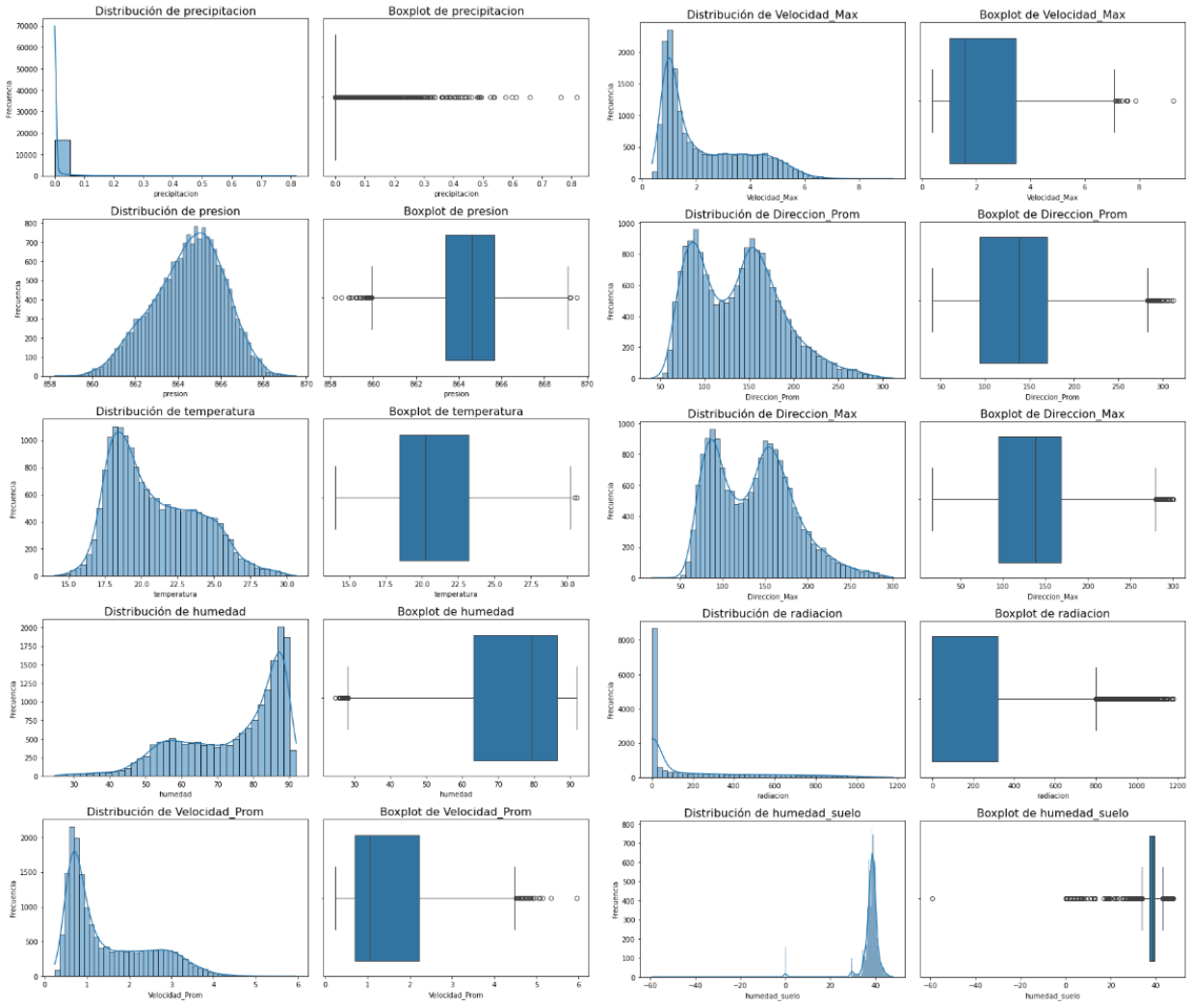


Figura 4. Distribución y boxplot para cada variable ambiental

Para detectar outliers en este punto se propusieron dos métodos (IQR y z-score) pero se optó por no usarlos debido a que estas variables ambientales presentan datos naturalmente sesgados, dispersos y asimétricos, por lo que la mayoría serían identificados como *outliers*. En el caso del z-score, este

método es definido para datos que tienden a distribuirse normalmente, situación que no sucede con la mayoría de las distribuciones. Mientras que el método del Rango Intercuartílico (IQR), aunque es no paramétrico, es menos efectivo para datos asimétricos o dispersos. Los métodos mencionados se describen en el *Anexo 1. Detección de outliers***.**

Preparación de los datos: Limpieza, filtrado y rellenado

Inicialmente se identificaron los datos faltantes por variable. Para esto se realizó un mapa de calor de datos faltantes con respecto a su temporalidad. En la Figura 5 se muestran los resultados de esta consulta para las diez (10) variables.

Se observa que ocho (8) variables presentan la misma cantidad de información ausente para un mismo periodo. Mientras que dos (2) variables (incluyendo la variable respuesta) muestran gaps de información en diversos periodos y solo algunos coinciden.

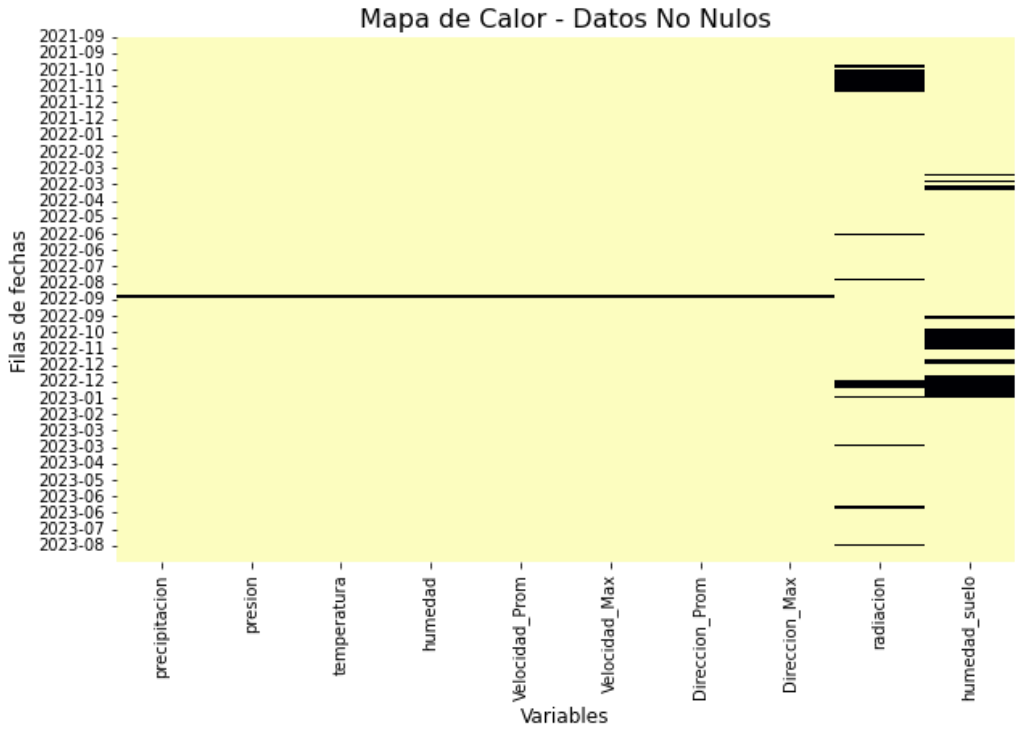


Figura 5. Mapa de calor que indica en color negro los valores nulos de las variables

Con el propósito de subsanar estos valores ausentes se propusieron diversas técnicas de inputado de datos, dependiendo de la cantidad de datos faltantes y la importancia de estas, las cuales se describen en la Tabla 4.

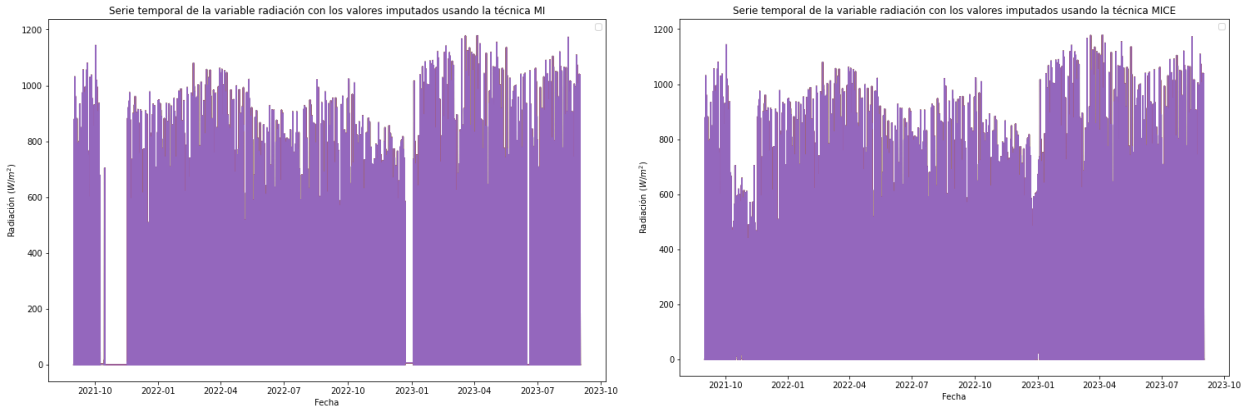
La primera técnica de tipo *baseline* por su sencillez, es *Median Imputation* (MI), la cual toma la mediana de cada variable y la usa como reemplazo de los valores faltantes. Esta se aplicó solo para las ocho (8) variables como se especifica en la tabla, pues estas tenían pocos datos para inputar.

La segunda técnica denominada *Multiple Imputation by Chained Equations* (MICE) fue implementada para la variable radiación. Este método utiliza otras variables y registros del conjunto de datos para predecir los valores que faltan[18].

Tabla 4. Lista de metodologías implementadas para imputación de datos por variable

Tipo	Nombre	Variables usadas									
		Prec	P	T	H	V. Prom.	V. Máx.	Dir. Prom	Dir. Máx.	Rad	CH
Baseline	Median Imputation (MI)	100	100	100	100	100	100	100	100		
Multivariado	Multiple Imputation by Chained Equations (MICE)									1,422	
Regresión	Linear regression										2,130

Cabe señalar que por principio de parsimonia se pretendía aplicar el método MI para la variable radiación. Sin embargo, como puede observarse en la Figura 6(a) los valores inputados mediante este método son inconsistentes con el comportamiento típico de la variable, ya que solo asigna un valor fijo para periodos de tiempo importantes, cuando esta variable presenta cambios en sus valores con frecuencias de horas. Si bien con MICE no se observa esta situación perfectamente subsanada, sí es un método más eficaz para inputar datos de este tipo de variables, tal como se observa en Figura 6(b).



(a) (b)

Figura 6. Serie temporal con todos los valores inputados de la variable radiación (W/m^2) usando los métodos MI (a) y MICE (b)

[18] KIM, Yohan, *et al*. Strategies for imputation of high-resolution environmental data in clinical randomized controlled trials. International Journal of Environmental Research and Public Health, 2022, vol. 19, no 3, p. 1307.

Para la variable respuesta se implementó una regresión lineal, la cual se entrenó con los datos conocidos para predecir los valores faltantes, los cuales se concatenaron posteriormente con los datos disponibles y así se obtuvo el dataset con todos los registros no nulos. Esta técnica fue elegida como un método más sofisticado de imputación que los anteriores.

Es importante aclarar que se analizó la matriz de correlación entre las diez (10) variables, presentada en la Figura 7, para identificar colinealidad antes de ejecutar la regresión lineal. De esta manera, se observó que las variables V. Prom y V. Máx. tienen una correlación de 0.99. De igual forma con las variables Dir. Prom y Dir. Máx. Estos valores indican que existe redundancia y por tanto se decidió eliminar del análisis las variables V. Máx y Dir. Máx. En ambos casos esta decisión fue tomada dado que se considera el valor promedio como un parámetro que representa mejor el comportamiento de la variable. Mientras que para el resto de las variables, aunque se presentaron pares con alta colinealidad, se decidió conservarlas debido a que por entendimiento del problema, se consideran relevantes para explicar el contenido de humedad del suelo.

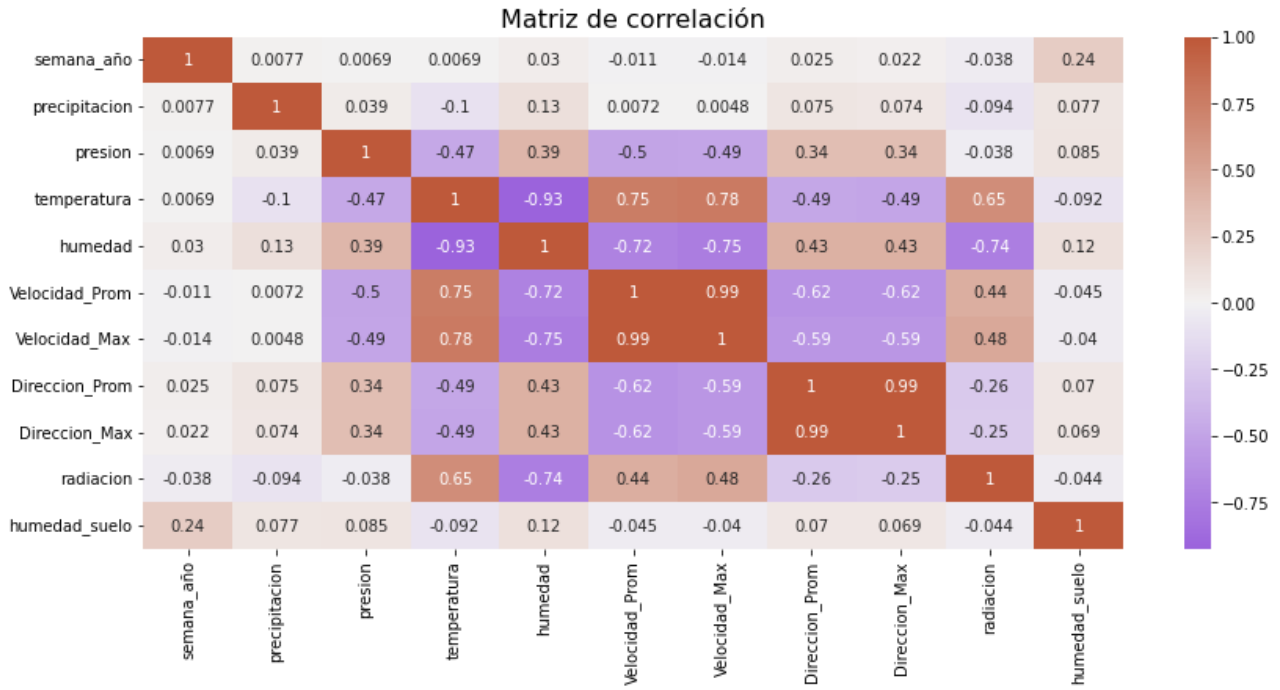


Figura 7. Matriz de correlación entre las variables del dataset

Transformación de los datos

Dadas las distribuciones evidenciadas en la Figura 4, donde se nota que de las variables seleccionadas para el modelo, solamente la presión y la humedad del suelo parecen tener una distribución normal. Al correr el modelo con las variables en sus valores originales, sin ser transformados, se obtuvo que los errores o residuales no presentaban un comportamiento normal, lo que es tomado como supuesto en la regresión lineal para poder hacer las pruebas de hipótesis respecto a los coeficientes de interacción de las variables. Por lo anterior, se decidió realizar la transformación de los datos bajo el

método Box-Cox, el cual busca corregir la asimetría (skewness) de una variable, mejorando sesgos en la distribución, ayudando a corregir la no linealidad en la relación, en caso de presentarse y también a normalizar la distribución de la variable respuesta y mejorar el ajuste del modelo. Así, esta transformación se define por la siguiente ecuación:

$$\text{Box-Cox} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & si\ \lambda \neq 0 \\ log\ (x) & si\ \lambda = 0 \end{cases}$$

La transformación por el método Box-Cox se realiza siguiendo los pasos:

1. Verificación de Positividad: La transformación Box-Cox solo se puede aplicar a datos positivos. Por lo tanto, antes de aplicarse, se verifica que todos los valores de los datos de entrada sean mayores que cero. En caso de que haya valores igual a cero, o menores[19], se aplica la transformación Yeo-Johnson, dada por:

$$\text{Yeo-Johnson} = \begin{cases} \dfrac{(x+1)^{\lambda} - 1}{\lambda} & si\ x \geq 0\ y\ \lambda \neq 0 \\ log(x+1) & si\ x \geq 0\ y\ \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda} - 1}{2-\lambda} & si\ x < 0\ y\ \lambda \neq 2 \\ -log(-x+1) & si\ x < 0\ y\ \lambda = 2 \end{cases}$$

2. Estimación de λ: Posteriormente se debe estimar el valor óptimo del parámetro λ que maximiza la verosimilitud de los datos transformados siendo normalmente distribuidos. Esto se realiza a través de un proceso de búsqueda que evalúa la función de verosimilitud en diferentes valores de λ.
3. Aplicación de la Transformación: Una vez estimado el valor óptimo de λ, se aplica la transformación a los datos usando las fórmulas mencionadas anteriormente.
4. Salida: Finalmente se obtiene como salida los datos transformados, el valor óptimo de λ utilizado para la transformación y el valor de la asimetría.

En la Tabla 5 se muestra la asimetría que presentan las variables antes de ejecutar la transformación. Los valores de asimetría más cercanos a cero indican mayor simétrica en una distribución. Si el valor es positivo indica que la cola derecha (valores altos) de la distribución es más larga o más pesada que la izquierda (como se puede ver para la precipitación y la radiación), mientras que un valor negativo indica que la cola izquierda (valores bajos) es más larga o más pesada que la derecha (como sucede con la variable humedad relativa y humedad del suelo). En el caso de la presión la simetría es alta, y para la temperatura y la velocidad promedio del viento, la asimetría es baja.

[19] En este caso en particular no se reportan valores negativos para las variables, pero sí ceros.

Tabla 5. Asimetrías de las variables antes y después de la transformación

Variables	Asimetría sin transformación	Asimetría con transformación
Humedad	-0.834	-0.330
Precipitación	10.277	2.528
Presión	nan	nan
Temperatura	0.575	0.049
Velocidad promedio del viento	0.927	0.055
Radiación	1.413	0.108
Humedad del suelo	-0.015	0.013

Escalado de los datos con min max

Dadas las diferentes escalas de medida y la presencia de variables con alta y baja varianza en estudio y las posibles implicaciones que esto podría representar para los modelos a probar: regresión lineal y knn (necesitando este último la medición de distancia entre las diferentes variables), se decidió hacer un escalamiento min max para normalizar los datos y dejarlos bajo la misma escala de medida (entre 0 y 1), con lo que se espera mejor ajuste del modelo. Esta es definida en la siguiente ecuación:

$$\text{min-max scaler} = \ \mathbf{0} \leq \frac{\boldsymbol{X - X_{min}}}{\boldsymbol{X_{max} - X_{min}}} \leq \mathbf{1}$$

Análisis descriptivo de los datos transformados

Una vez realizada la imputación y preparación de los datos, se procede con el análisis de los datos transformados. En la Tabla 6 se presenta un resumen de los principales estadísticos de tendencia central, dispersión y localización para las variables meteorológicas.

Tabla 6. Estadísticos descriptivos de variables meteorológicas después de la transformación y escalado

	Humedad Relativa	Precipitación	Presión	Temperatura	Velocidad Promedio	Radiación	Humedad Del Suelo
Count	17521	17521	17521	17521	17521	17521	17521
Mean	0.572	0.108	0.478	0.600	0.579	0.414	0.498
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.306	0.000	0.374	0.464	0.422	0.000	0.404
50%	0.634	0.000	0.483	0.585	0.548	0.451	0.505
75%	0.828	0.000	0.582	0.744	0.765	0.849	0.590
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Std	0.277	0.278	0.147	0.170	0.193	0.413	0.159

De la tabla anterior, en la que se presentan los datos ya transformados y escalados, se puede ver que las variables presión, temperatura, velocidad y humedad del suelo son las que presentan un comportamiento con menos variación que se puede ver en su desviación estándar, que respecto a la media no representa valores altos; esto mismo se puede ver cuando se compara los valores de la media y la mediana, que no presentan diferencias significativas entre sí para cada una de estas variables. Lo anterior hace sentido dada la naturaleza de estas variables (por ejemplo, la presión que depende principalmente de la altura), y por supuesto del sensor, que no se espera presente cambios abruptos en su medición.

Por otro lado, las variables humedad relativa, precipitación y radiación, parecen presentar una mayor variación, ocasionando mayor diferencia entre la media y la mediana, y que se evidencia también en el valor del cuartil 3. En particular con la precipitación se observa que todos los valores están concentrados debajo de cuartil 1, indicando unos valores muy pequeños en la medida. Considerar una transformación de estos datos no es viable se considera parte del comportamiento natural de la variable. Las variables humedad relativa y radiación, si bien presentan mayor variación que las demás, no parecen tener datos atípicos alarmantes.

A continuación, se presentan los histogramas de frecuencia junto con los box-plot para mejor ilustración del comportamiento de las variables.

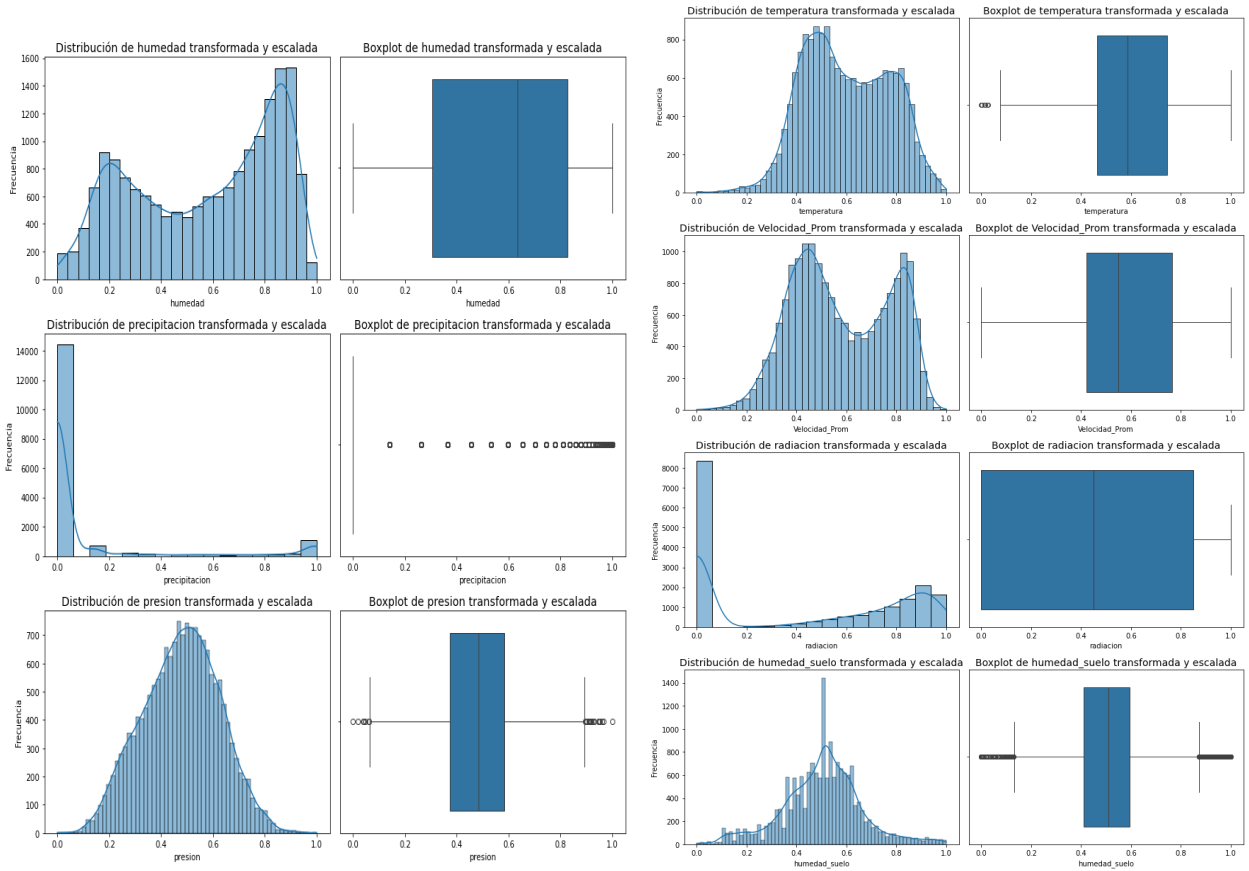


Figura 8 Histogramas y boxplot de variables con datos escalados y transformados

En el reporte descriptivo adjunto en *Anexo 2: Reporte descriptivo de los datos*, y en la Figura 8, se nota que para la humedad relativa (HR), la temperatura, y la velocidad promedio del viento, sus distribuciones mejoraron considerablemente el sesgo, lo que las lleva a una forma más adecuada para su análisis de correlación y su uso en el modelo de regresión. Sin embargo, es importante destacar que para las variables HR y temperatura, luego de la transformación, los datos tienden a presentar dos modas: en el caso de la HR, puede deberse a los cambios que se dan en diferentes momentos del año (aumenta en temporadas de lluvia normalmente presentes entre marzo y junio, y septiembre hasta diciembre, siendo mayor en la época lluviosa del segundo semestre y menor en la época seca de julio[20]). Sin embargo, en los gráficos de dispersión presentes en el reporte adjunto, se encuentra que, a pesar de la presencia de estas dos modas, la variable de interés, no parece verse alterada frente a estos. Podría decirse que en los valores altos de HR, se nota un leve cambio en la humedad del suelo, mostrando cambios marginales, aunque no drásticos.

Para la temperatura se manifiesta el mismo fenómeno que la humedad con respecto a las estaciones climatológicas, adicional a los cambios diarios que se presentan con la radiación solar. Al observar la humedad del suelo frente a la temperatura, se ve que los cambios en la humedad del suelo se dan especialmente para valores bajos de temperatura.

Por su parte, para las variables presión y humedad del suelo, se observa que presentan una distribución aproximadamente normal, lo cual las hace adecuadas para su análisis de correlación y adicional, su interacción en el análisis de regresión. En particular para la presión, al analizar la dispersión del archivo adjunto, no se infieren cambios en la humedad del suelo dados los valores que va tomando la presión.

Respecto a la velocidad promedio del viento, en principio no es evidente una relación clara con la humedad del suelo, pero se logra evidenciar cambios en los valores de la humedad del suelo para valores bajos de la velocidad promedio. Dado que Medellín es una zona relativamente húmeda, a pesar de que pueda haber alguna interacción, no se espera un efecto considerable, como si podría darse en zonas más áridas.[21]

Finalmente, respecto a la dispersión de los datos, se tiene que las variables temperatura, humedad, del suelo, y presión, presentan menor dispersión que las demás variables. Esto puede verse tanto en los histogramas como en los boxplot presentes en la Figura 8.

Respecto a las variables radiación y precipitación, dados los valores presentes en los boxplot de la Figura 8Figura 4, se decidió binarizarlas para reducir la dispersión de los datos y de esta forma obtener más conocimiento de su comportamiento. En la Figura 9 se presenta el resultado de este ejercicio, donde se nota especialmente para la precipitación, que hay una diferencia para la cantidad

[20] IDEAM. Características climatológicas de ciudades principales y municipios turísticos [en línea]. [Consultado el 7, junio, 2024]. Disponible en Internet: http://www.ideam.gov.co/documents/21021/418894/Características+de+Ciudades+Principales+y+Municipios+Turísticos.pdf/c3ca90c8-1072-434a-a235-91baee8c73fc.

[21] JARAMILLO, Daniel F. Introducción a la ciencia del suelo. 2002.

de datos con etiqueta de precipitación y no precipitación por el comportamiento natural de la variable explicado en la Sección Descripción de los datos. Mientras que para la variable radiación, se tiene un comportamiento más uniforme entre etiquetas, pero cuenta levemente con más registros en momentos de luz, lo cual es esperado debido al comportamiento de la luz solar en el trópico.

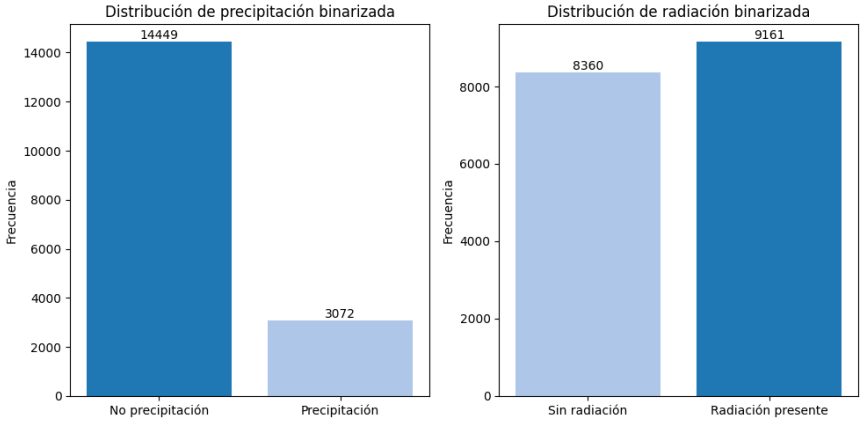


Figura 9 Variables precipitación y radiación binarizadas

A continuación, se presenta la matriz de correlación, con los datos transformados y escalados.

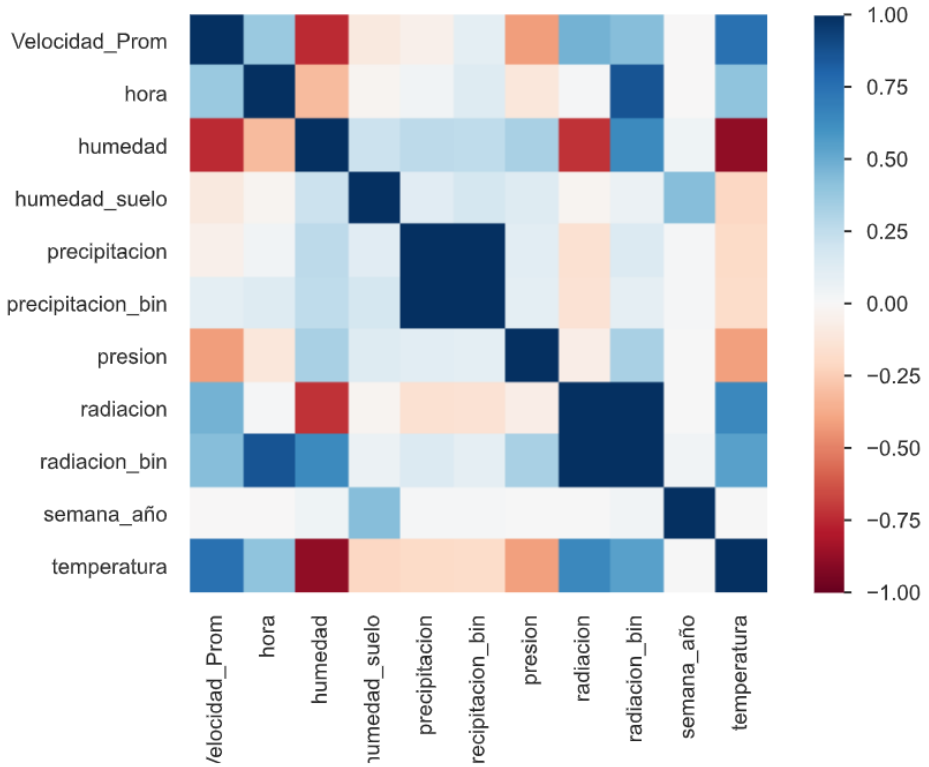


Figura 10 Matriz de correlación entre las variables, con datos escalados y transformados

De acuerdo a la matriz anterior, se nota que luego de la transformación, aunque leve, es más clara la correlación de las variables, temperatura, humedad y velocidad promedio con la humedad del suelo, así mismo, se ve que la variable temporal para la que se evidencian más cambios en la humedad del suelo, es la variable semana.

c. Modelación de los datos

Con el fin de estimar las relaciones entre la variable respuesta y las variables predictoras, teniendo en cuenta la distribución de los datos y su naturaleza continua, se propone realizar una serie de modelos de regresión, con dos tipos de conjuntos de datos y dos tipos de métodos de regresión: paramétrico y no paramétrico, como se especifica en la Tabla 7. Modelos Tabla 7.

Tabla 7. Modelos propuestos

Conjunto de datos	Método de regresión	Nombre del modelo
Variables ambientales + hora como variable categórica	Regresión lineal	RL1
	KNN	KNN1
Variables ambientales SIN hora	Regresión Lineal	RL2
	KNN	KNN2

El flujo de trabajo para la ejecución del modelo se muestra en la Figura 11 y se desglosa en las secciones siguientes.

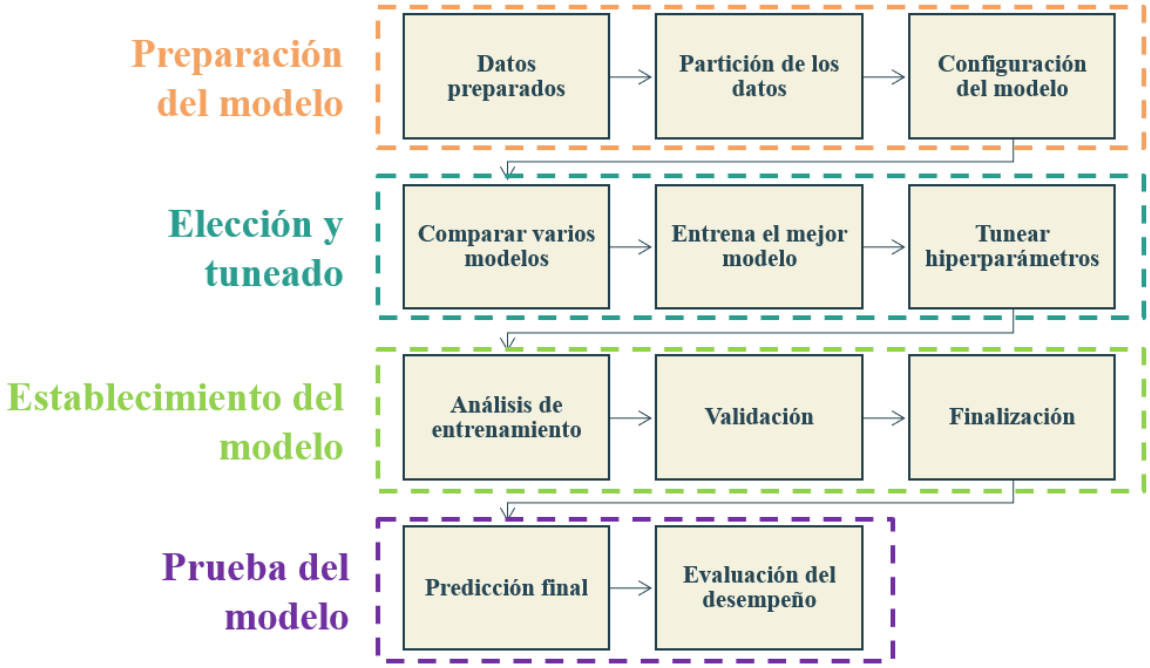


Figura 11 Metodología para la modelación de los datos

Preparación del modelo

En esta fase entrena el modelo con los datos transformados y escalados. Para entrenar el modelo se separaron los datos en entrenamiento y testeo (o datos no vistos). Para esto, se tomó la serie temporal de la base de datos preparada y se seleccionó el mes de noviembre de 2021 para el testeo, ya que es un mes sin eventos de lluvia o sequía pronunciados, por lo que considera que representa un comportamiento promedio. De forma que los datos de entrenamiento se determinaron como todos los demás registros excepto los de este mes.

En la configuración del modelo se le ingresa la base de datos, se define cuál será la variable respuesta, y si hay variables categóricas o si hay variables para ignorar en el entrenamiento del modelo. Acá se parte los datos de entrenamiento en entrenamiento (70%) y validación (%30). Además se separa la variable respuesta de las variables predictoras, quedando 4 sets de datos: X entrenamiento, Y entrenamiento, X validación, Y validación (estos últimos también llamados hold-out sample o muestras de reserva).

Elección y tuneado

En esta etapa inicia entrenando varios modelos a la vez sobre el conjunto de prueba, y calculando las métricas de desempeño R^2, MAPE y el tiempo de entrenamiento en segundos de cada modelo. Esto permite tener un criterio para elegir qué modelo puede ser mejor.

Una vez identificado el mejor modelo, se entrena solo ese modelo con una validación cruzada de 8 folders sobre el conjunto de entrenamiento. Se continua ajustando los hiperparámetros, que en el caso de la regresión lineal son el intercepto y la pendiente (los betas), de forma que el error entre el valor predicho y el real se minimice, mientras que en caso del KNN se optimiza el valor de k y la métrica de distancia. La mejor combinación de hiperparámetros se puede obtener probando todos los posibles hiperparámetros, uno por uno, o probando de forma aleatoria. Este primer método es llamado *GridSearch* y fue el escogido.

Establecimiento del modelo

Ya con el modelo entrenado y los hiperparámetros tuneados, se analiza la calidad del entrenamiento mediante el análisis de los residuales, el error de predicción y la importancia de características solo para LR1 y LR2.

Si se observa que el modelo tiene un buen ajuste con los datos de entrenamiento, se corre para predecir los datos de validación o la ‘muestra de reserva’. Finalmente se ajusta el modelo a los datos completos (entrenamiento y validación juntos) y se congelan los hiperparámetros, en otras palabras, se guardan los hiperparámetros calculados.

Prueba del modelo

Para la prueba final del modelo o para usarlo en producción, se predice el conjunto de testeo o datos no vistos de noviembre de 2021 y por último se evalúa el desempeño del modelo con las métricas elegidas.

A continuación, se presentan los resultados obtenidos para los hiperparámetros de los modelos así como sus métricas de desempeño.

Ejecución de los modelos

Al ejecutar los modelos con los datos escalados se obtienen los siguientes resultados para cada uno de los modelos:

Tabla 8. Métricas de los modelos comparados

Sigla	Model	R^2	MAPE	TT (Sec)
knn	**K Neighbors Regressor**	**0.2606**	**0.3684**	**0.0800**
lr	**Linear Regression**	**0.2362**	**0.4016**	**0.1000**
ridge	Ridge Regression	0.2362	0.4018	0.1000
lasso	Lasso Regression	-0.0001	0.4733	0.0900
en	Elastic Net	-0.0001	0.4733	0.1100

Entrenamiento de un modelo KNN, tomando la hora como variable categórica

Se decide evaluar solamente los modelos KNN y Regresión Lineal; ya que en la tabla anterior con los datos de entrenamiento se notó que en principio el KNN presenta un mejor resultado tanto en la métrica de R2, relacionada al porcentaje de varianza de la variable respuesta explicado por el modelo, como en el MAPE, relacionado al % de error en la predicción respecto al valor real de la variable respuesta. Por lo que se decide entrenar el modelo usando validación cruzada y posteriormente evaluar su desempeño con los datos de prueba, es decir, los datos no vistos. A continuación, las métricas de desempeño luego del entrenamiento con validación cruzada y evaluado con los datos de prueba.

Tabla 9 Métricas de desempeño modelo KNN con datos de prueba después del tuneado de hiperparámetros con validación cruzada

	Model	R2	MAPE
0	**KNN**	-0.7796	0.1186

De la tabla anterior se nota que, a pesar de que con los datos de entrenamiento el modelo mostró un R2 positivo y que aparentemente lograba explicar el 26% de la varianza de la variable respuesta, al entregarle los valores no vistos, el modelo como se planteó está mostrando un R2 negativo, del -0.7796, lo que indica que este no logra explicar la variabilidad presente en la variable respuesta, lo anterior puede darse porque el modelo se aprendió exactamente el comportamiento de los datos de entrenamiento y los datos de prueba son diferentes de estos en un punto en el que el modelo no logra

hacer la predicción, también es posible que la presencia de datos atípicos en las diferentes variables ingresadas al modelo no hayan permitido un entrenamiento correcto.

Entrenamiento de un modelo de regresión lineal usando la hora como variable categórica.

Para el modelo de regresión Lineal también se realizó el entrenamiento con validación cruzada, a continuación se muestran las métricas para evaluación del desempeño con los datos de entrenamiento.

Tabla 10 Métricas de desempeño del modelo de Regresión lineal con los datos de entrenamiento.

	R2	MAPE
Mean	0.2365	0.3882

Con los datos de entrenamiento, se obtuvo un Mape del 38.83% y un R2 de 23.65%.
A pesar de que es un R2 bajo, es aceptable teniendo en cuenta que se trata de variables metereológicas y que es relativamente corto el tiempo de entrenamiento.
En la siguiente figura se muestra la distribución de los residuales luego de la predicción realizada con el modelo tuneado, en donde se nota que presentan una distribución normal, lo anterior es importante porque es uno de los supuestos que deben cumplirse para que las pruebas de hipótesis respecto a los coeficientes tengan validez.

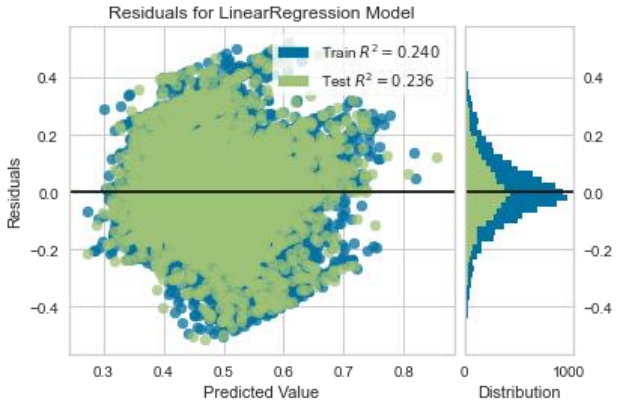


Figura 12 Distribución de los residuales luego de la predicción.

Una vez validada la normalidad en los errores luego de la predicción, se presenta a continuación los resultados obtenidos del modelo y las pruebas de hipótesis para cada uno de los beta. Es importante aclarar que los resultados anteriores fueron obtenidos mediante el modelo de regresión entrenado usando la libería pycaret, sin embargo, esta no permite realizar la prueba de hipótesis para cada uno de los betas, en consecuencia lo que se hizo, fue tomar el intercepto obtenido en el modelo, y entregar los mismos datos usados por pycaret en stats, librería que sí permite generar la prueba de hipótesis. Se sabe que no es exactamente el mismo modelo pero que es una aproximación a las relaciones ajustadas por el modelo. A continuación, en la Tabla 11 se presentan los resultados obtenidos.

Tabla 11 Resultados del modelo seleccionado

OLS Regression Results						
Dep. Variable:	humedad_suelo			**R-squared:**		0.194
Model:	OLS			**Adj. R-squared:**		0.194
Method:	Least Squares			**F-statistic:**		354.8
Date:	Fri, 16 Jun 2024			**Prob (F-statistic):**		0.00
Time:	0,775671296			**Log-Likelihood:**		6192.9
No. Observations:	11776			**AIC:**		-1,24E+07
Df Residuals:	11767			**BIC:**		-1,23E+07
Df Model:	8					
Covariance Type:	nonrobust					
coef	**std**	**Err***	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.4202	0.021	20.173	0.000	0.379	0.461
semana_año	0.1720	0.005	37.543	0.000	0.163	0.181
humedad	0.0664	0.013	5.092	0.000	0.041	0.092
presion	0.0449	0.010	4.372	0.000	0.025	0.065
temperatura	-0.3311	0.020	-16.841	0.000	-0.370	-0.293
Velocidad_Prom	0.1323	0.011	11.585	0.000	0.110	0.155
precipitacion_bin	0.0225	0.004	6.088	0.000	0.015	0.030
radiacion_bin	0.0609	0.004	17.138	0.000	0.054	0.068
hora	0.0018	0.000	8.188	0.000	0.001	0.002
Omnibus:		276.974		**Durbin-Watson:**		1.998

*El error estándar supone que la matriz de covarianza de los errores está correctamente especificada.

De acuerdo a la tabla anterior, en la que se ven los estadísticos de prueba y coeficientes de cada uno de los betas para cada una de las variables independientes respectivas, se infiere que, dado que todos los coeficientes tienen valores p muy pequeños ($p < 0.05$), hay suficiente evidencia para rechazar la hipótesis nula relacionada a que los valores de los beta no son diferentes de cero, lo que indica que todas las variables independientes en el modelo son estadísticamente significativas para predecir la humedad del suelo. Por otro lado, al ver el rango, estos no contienen el valor cero. Respecto al R2, a pesar de que este es relativamente bajo, el modelo puede proporcionar información útil sobre la relación entre las variables independientes y la humedad del suelo. Sin embargo, es importante mencionar que el porcentaje de varianza de la variable (humedad del suelo) explicada por el modelo es poca, por ende, si se quiere mejorar su explicabilidad o predicción, sería necesario añadir más variables relevantes o si se pueden utilizar técnicas más avanzadas para mejorar la capacidad predictiva del modelo.

A continuación, en la Tabla 13 se muestran las métricas de desempeño con los datos de prueba.

Tabla 13 12 Métricas de desempeño con los datos de prueba para el modelo de regresión lineal.

	MAPE	R2
Mean	0.021	-0,639

De acuerdo al resultado anterior, a pesar de que en los datos de entrenamiento se obtenía un R2 positivo y aceptable en el campo de estudio, luego de correrlo con los datos de prueba, se obtiene un R2 negativo, a pesar de que es levemente mejor que el obtenido mediante KNN, es también insuficiente para hacer afirmaciones sobre un modelo que sea capaz de explicar la varianza de la humedad del suelo (variable respuesta), dado el comportamiento de las demás variables consideradas. Lo anterior puede deberse a la alta variabilidad presente y evidenciada en la etapa de descripción de los datos, en la que se vio que a excepción de la variable presión, las demás presentan alta desviación y valores extremos, y hace parte de su naturaleza, mientras que para la humedad del suelo se notó que los datos tienen a estar menos dispersos, es decir, presentan menos variación. Adicional a lo anterior, como se mencionó en el análisis de resultados del modelo KNN, dada la alta variabilidad presente en los datos, en la complejidad del comportamiento de estas ya que pueden ser afectadas por varios factores naturales, 2 años de entrenamiento pueden considerarse pocos para entrenar un modelo que logre explicar la varianza presente en la variable respuesta y su relación con las variables explicativas.

Tabla 13. Métricas resumidas de todos los modelos.

Conjunto de datos	Nombre del modelo	Métricas en entrenamiento	Métricas en validación	Métricas en prueba	Métrica des-escalada
Variables ambientales + hora como variable categórica	RL1	R2: 0.236 MAPE: 0.388	R2: 0.236 MAPE: 0.401	R2: -0.633 MAPE: 0.119	R2: -0.639 MAPE: 0.021
	KNN1	R2: 0.290 MAPE: 0.340	R2. 0.296 MAPE: 0.354	R2: -0.803 MAPE: 0.118	
Variables ambientales SIN hora	RL2	R2: 0.188 MAPE: 0.402	R2: 0.185 MAPE: 0.417		
	KNN2	R2: 0.206 MAPE: 0.309	R2: 0.188 MAPE: 0.332		

4. Despliegue

Una vez finalizado el proceso de modelado, obtenemos la predicción de la humedad del suelo del mes propuesto, así como los datos de entrenamiento para el mismo mes, en archivos de formato Excel como salida. Siendo esto insumos para la visualización o reporte grafico de los resultados, fuente importante para el análisis y monitoreo constante de los datos.

La herramienta de visualización seleccionada para esta fase es Power BI, debido a que en la industria de la analítica de datos es ampliamente reconocida por su robustez y respaldo a la hora de conectar, modelar y comparar información.

En el cuadrante de Garner del 2023, Microsoft con Power BI se encuentra como líder absoluto sobre Plataformas Analíticas y de Business Intelligence (ABI).

Figura 13 Cuadrante de Gartner sobre plataformas Analíticas y de Business Intelligence (ABI).

a. Storytelling de los datos

El tablero de Power BI creado consta de tres páginas: portada, modelo y descripción. En la página modelo se presenta el comparativo de los datos de humedad del suelo reales VS los predichos por el modelo. Mientras que en la página descripción se grafica los datos históricos de humedad del suelo respecto a cada una de las variables meteorológicas.

Pronóstico

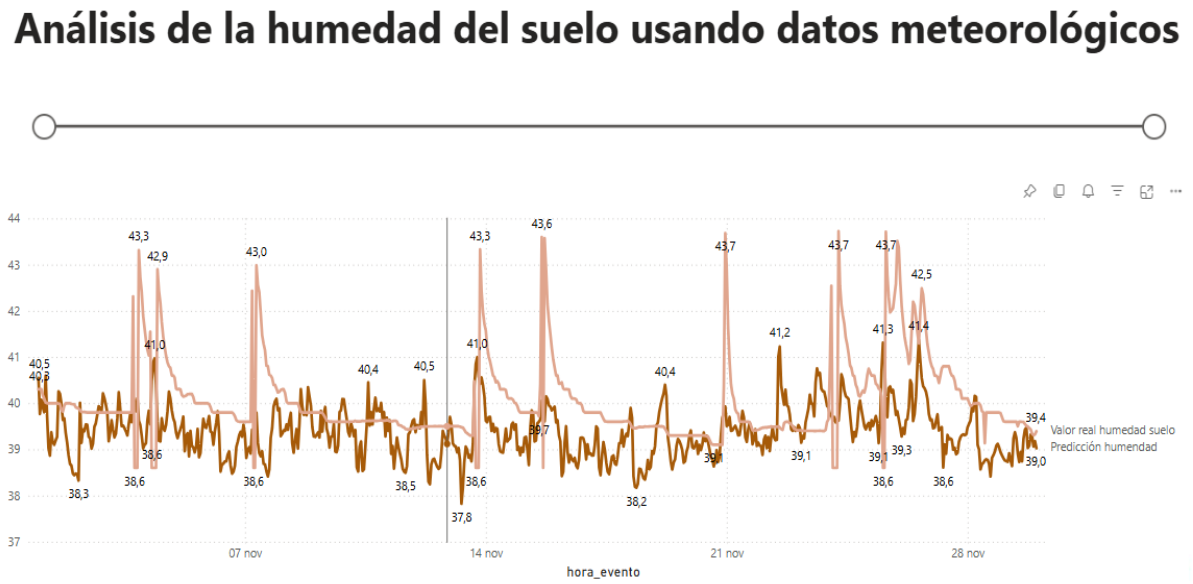


Figura 14 Pronóstico vs datos reales de la variable humedad del suelo, tomado de la página 'modelo' del tablero en Power BI.

Se evidencia que el modelo trata de simular el comportamiento de la variable humedad suelo, pero de una forma conservadora, viendo que en las altas estaciones falla en algunos casos dado que subestima su predicción, pero en general se observa un comportamiento símil.

Análisis de la humedad del suelo con otras variables meteorológicas

Al visualizar el comparativo de las series temporales entre la humedad del suelo y el resto de variables meteorológicas, es evidente sus propiedades estacionales, presentando múltiples picos y valles en una frecuencia aparentemente regular. Es sobre todo evidente al comparar la precipitación con la humedad del suelo que algunos picos son compartidos. Como trabajo a futuro, el análisis de estas variables mediante series de tiempo sería la forma más efectiva de obtener conocimiento de estos datos.

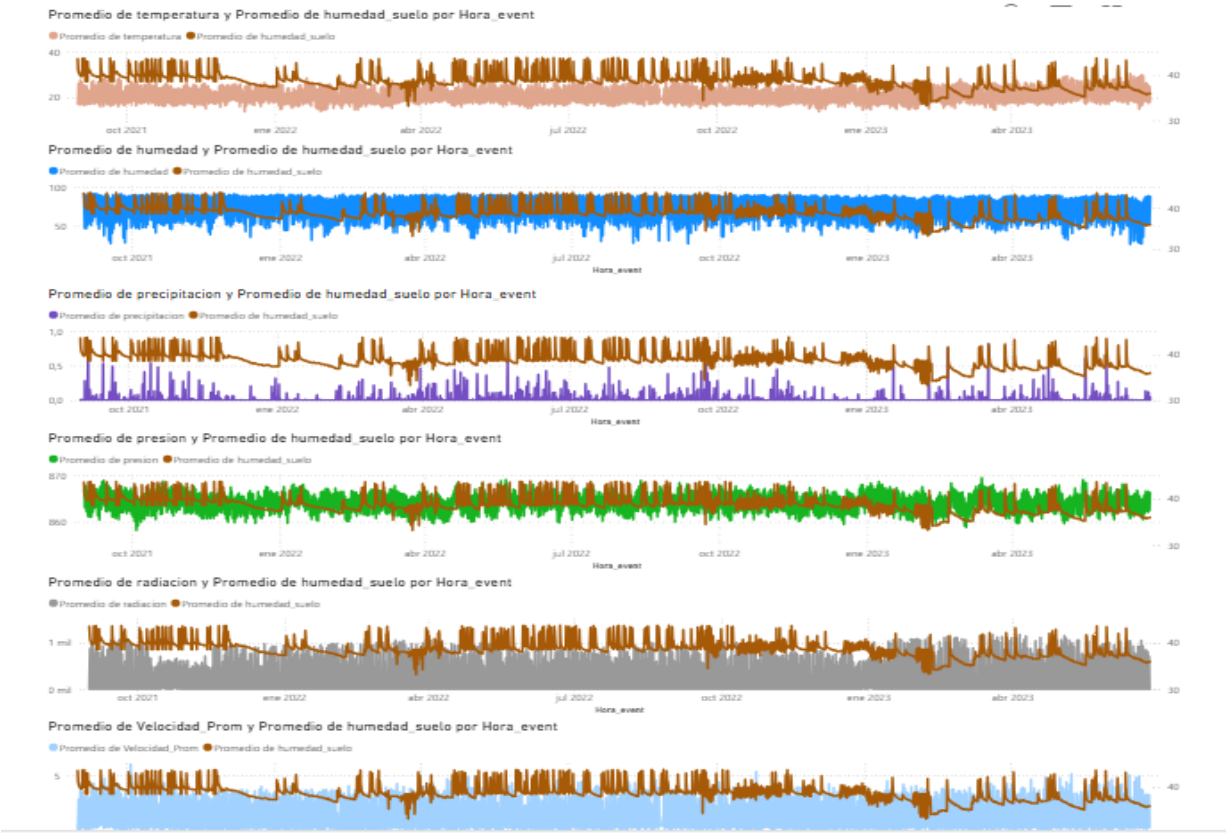


Figura 15 Series de tiempo del promedio de humedad en el suelo por hora vs las series de tiempo promedio de temperatura, humedad del aire, precipitación, presión atmosférica radiación y velocidad promedio del viento. Tomado de la página 'descripción' del tablero en Power BI.

Por otra parte, la variable presión atmosférica y humedad relativa del aire en su naturaleza son altamente estacionales, con picos marcados a una frecuencia constante, a diferencia de la variable humedad del suelo que muestra una tendencia de menor frecuencia en la ocurrencia de sus picos.

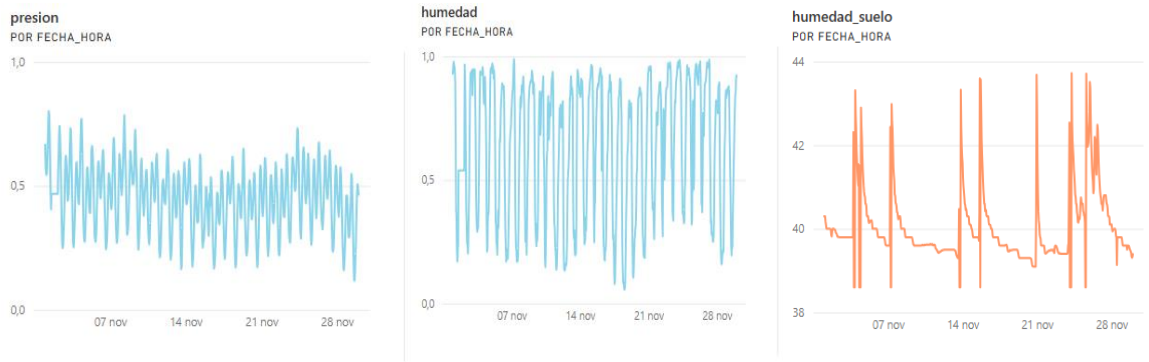


Figura 16 Series de tiempo del promedio presión atmosférica, humedad del aire y de humedad en el suelo por hora para el mes de noviembre de 2021

Se encontraron relaciones relevantes entre algunas variables ambientales del conjunto de datos, como la de la variable humedad del aire con la temperatura, en la que existe una correlación marcada.

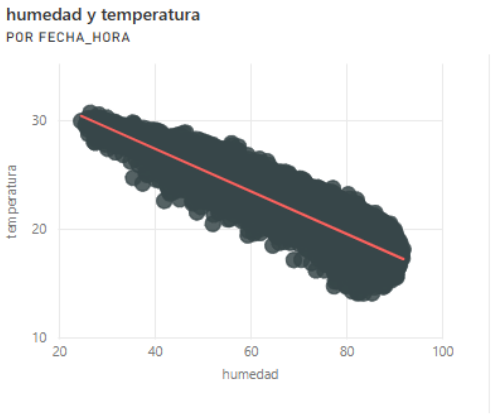


Figura 17 Correlaciones relevantes entre las variables del conjunto de datos

La predicción del modelo como resultado arroja a diferencia de lo real una alta estacionalidad

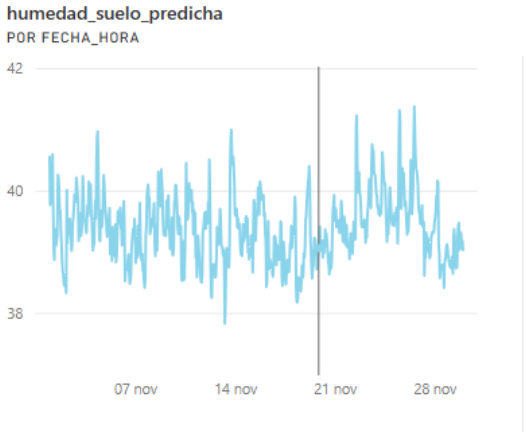


Figura 18 Serie de tiempo predicha del modelo seleccionado

Actualmente el reporte de los datos está en área de trabajo privado y se puede acceder a ellos con aprobación del administrador del reporte, solicitando permisos de lectura de la URL:

https://app.powerbi.com/links/VILsPvZqtM?ctid=99f7b55e-9cbe-467b-8143-919782918afb&pbi_source=linkShare

Es posible crear un grupo de Teams donde se puedan otorgar permisos de lectura a sus participantes de manera masiva o publicar el reporte como objeto visual en una aplicación WEB.

c. Posible caso de uso

A nivel mundial la Red Internacional de Humedad del Suelo (ISMN por sus siglas en inglés) dispone de un geoportal visualizado en la Figura 19, donde se puede graficar el histórico de datos de humedad del suelo medida con los sensores de dicha red, así como datos de otras variables meteorológicas disponibles.

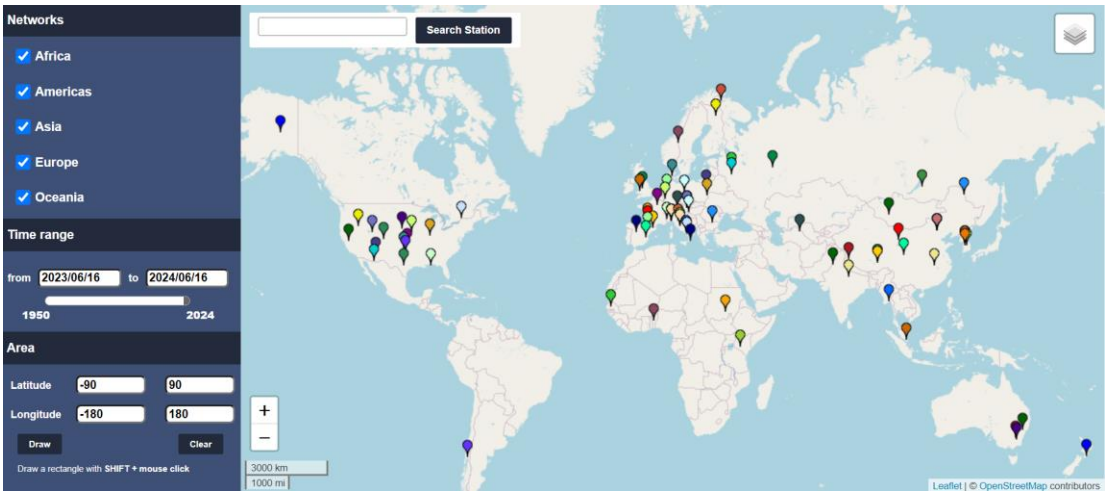


Figura 19. Geoportal de ISMN

Para Colombia solo se cuenta con una estación en la Guajira, con histórico desde 2013 hasta 2016 de la humedad del suelo, precipitación y temperatura del aire, evidenciado en la Figura 20.

Siendo la humedad del suelo una variable de interés para muchos sectores, el posible caso de estudio consiste en amplificar esta red en Colombia, donde se puedan visualizar estimados de la humedad del suelo para estaciones donde solo se midan variables meteorológicas. Es por esto que los resultados obtenidos con este proyecto, se constituyen como un primer paso para lograr este acometido.

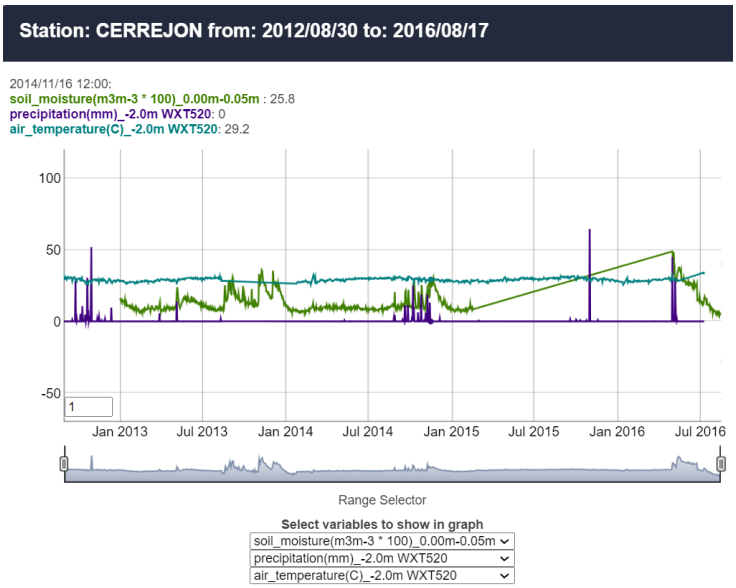


Figura 20. Histórico de datos para la estación en Colombia.

5. Conclusiones.

- La variable hora, a pesar de ser generada inicialmente para tener control de la temporalidad en el modelo como variable independiente, resultó ser relevante para el ajuste del modelo, lo anterior puede dar lugar a inferir, que, a posiblemente existe variables no incluidas en el modelo que dependan del tiempo y sean relevantes para la humedad del suelo.
- De las variables estudiadas, solo la presión atmosférica presentó un comportamiento normal, esto es debido a que a nivel local la presión atmosférica es relativamente constante con pequeñas variaciones a lo largo del año.
- Al realizar la transformación y escalamiento de los datos para llevarlos a distribuciones menos sesgadas y escalas más comparables, el modelo presentó mejor ajuste, lo que entre otras razones, corrobora la teoría de que los datos escalados permiten mejores comparaciones para evaluar relaciones entre las diferentes variables.
- Con el modelo de regresión lineal se obtuvo un R^2 relativamente bajo, esto es, menor al 30%, si bien, lo anterior significa que con el modelo desarrollado, se logra explicar un porcentaje de la variabilidad dada en la variable respuesta, humedad del suelo, para poder tener un mejor entendimiento de la variabilidad de *y*, sería necesario estudiar variables adicionales, o evaluar otros modelos con métodos más complejos que permitan modelar relaciones no lineales, entre las variables en estudio.
- De acuerdo a los resultados obtenidos en las pruebas de hipótesis respecto a la importancia de los betas para el modelo de regresión lineal entrenado con validación cruzada, hay suficiente evidencia estadística para afirmar que todas las variables independientes en el modelo son estadísticamente significativas para predecir la humedad del suelo. Respecto al R^2, a pesar de que este es relativamente bajo, el modelo puede proporcionar información útil sobre la relación entre las variables independientes y la humedad del suelo. Sin embargo, es importante mencionar que el porcentaje de varianza de la variable (humedad del suelo) explicada por el modelo es poca, por ende, si se quiere mejorar su explicabilidad o predicción, sería necesario añadir más variables relevantes o si se pueden utilizar técnicas más avanzadas para mejorar la capacidad predictiva del modelo. Adicional a lo anterior, este modelo funcionó para los datos vistos por el modelo, pero al entregarle los datos de prueba, quizá dada la alta variabilidad de las variables, el corto tiempo de entrenamiento, o la falta de variables adicionales, el modelo no tuvo un desempeño aceptable para realizar la predicción.
- Durante el análisis descriptivo de los datos, se encontró que las variables precipitación y radiación, representando su naturaleza, se comportan casi como una variable binaria y además sesgada; en el caso de la precipitación, la mayoría de las veces se encuentra en 0,pues es mayor la temporada seca a la temporada de lluvias; en contraste, para la radiación los valores em cero son casi los mismos que los registros que no lo son y por tanto es una variable binaria balanceada cuando se asigna cero a la ausencia de radiación y 1 cuando

el sensor detecta radiación. Al convertirlas en variables binarias presentaron mejor importancia en el modelo.

- Al entrenar el modelo KNN, que dada la distribución de la mayoría de las variables presentaban alta variabilidad, se obtuvo muy buen ajuste con el set de entrenamiento, sin embargo, al probarlo con el set de testeo, el desempeño del modelo fue insuficiente, pues su ajuste fue menor al 10%, con lo anterior, se evidenció una de las desventajas de este tipo de modelos no paramétricos, y es que se aprende tanto los datos de entrenamiento, que cuando se le muestran datos distintos, dado su sobreajuste, el modelo ya no tiene la capacidad de predecir correctamente los datos no vistos, especialmente con el conjunto de datos del presente estudio, para el que mes a mes se presentaba alta variabilidad y no necesariamente el conjunto de entrenamiento resultaría muy similar el conjunto de prueba.
- En el análisis de desempeño de los modelos, al hacer el tuneado para knn, se observó que a pesar de que la métrica para el MAPE mejoraba, el R^2 disminuía, lo anterior podría interpretarse como que el modelo entrenado está explicando menor porcentaje de variabilidad de la variable respuesta, pero con un error más bajo.
- De acuerdo con los expertos, para las variables meteorológicas, dada su complejidad, es importante tener un histórico suficiente de al menos 10 años, sin embargo, dada la dificultad de acceso a los datos sugirieron que podría ser aceptable tener al menos 1 año, en esta ocasión el análisis se realizó con 2 años de registro histórico. Al ser las variables meteorológicas un campo complejo de predecir, y que además, su comportamiento ha estado con alta variabilidad en los últimos años dados los efectos del cambio climático; una posible razón para enfrentarse a los problemas de ajuste de los modelos, pudo haber sido que el histórico de entrenamiento no fue suficiente y que dada la alta variabilidad presente en los datos, un solo ciclo, como predictor de un nuevo mes no sea suficiente para obtener un ajuste y una predicción lo suficientemente acertada; por otro lado, dada la alta variabilidad, pudo suceder también que el mes elegido, no fuese representativo frente al comportamiento que se había tenido en los datos o población de entrenamiento.
- La medición de la humedad del suelo es un tema aún experimental y complejo de modelar, por ende, los instrumentos de medición aún presentan oportunidades de mejora en su precisión, en esta etapa en que los datos registrados con relación al histórico ideal aún son pocos, se logra un aporte significativo desde el presente estudio es la imputación de los datos y la visualización de estos para poder monitorear esos momentos de falla del instrumento, es decir, en donde deja de capturar datos, así como la consistencia de los valores registrados. Adicional, a entender la alta varianza presente en los valores y la complejidad en el entendimiento de sus relaciones.

6. Referencias

ANELLO, Mirko, et al. Robust Statistical Processing of Long-Time Data Series to Estimate Soil Water Content. Mathematical Geosciences, 2024, vol. 56, no 1, p. 3-26.

BABAEIAN, Ebrahim, M. SADEGHI, S. B. JONES, C. MONTZKA, H. VEREECKEN, and M. TULLER. Ground, proximal, and satellite remote sensing of soil moisture. Reviews of Geophysics, 2019, vol. 57, no 2, p. 530-616., doi: 10.1029/2018RG000618.

Consultora tecnológica publica Cuadrante Mágico de Gartner sobre Plataformas Analíticas y de BI. Disponible en: https://blog.bismart.com/microsoft-power-bi-lider-cuadrante-magico-gartner-2023

EL MGHOUCHI, Youness. Solar energy modelling and forecasting using artificial neural networks: a review, a case study, and applications. En Artificial Neural Networks for Renewable Energy Systems and Real-World Applications. Academic Press, 2022. p. 113-147. 2.

FARAGO, T. Soil moisture content: Statistical estimation of its probability distribution. Journal of Applied Meteorology and Climatology, 1985, vol. 24, no 4, p. 371-376.

FENG, Yu, et al. Estimation of soil temperature from meteorological data using different machine learning models. Geoderma, 2019, vol. 338, p. 67-77. https://doi.org/10.1016/J.GEODERMA.2018.11.044.

HAYA, Pablo. La metodología CRISP-DM en ciencia de datos. Instituto de Ingeniería de Conocimiento. 2021. Accedido: may. 12, 2024 [En línea]. Disponible en: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

HUSSAIN, Mariam; SHARMIN, Nusrat; SHAFIUL, Sumayea Binte. Estimation of Soil Moisture with Meteorological Variables in Supervised Machine Learning Models. En 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2023. p. 1-6. https://doi: 10.1109/ECCE57851.2023.10101650.

JARAMILLO, Daniel F. Introducción a la ciencia del suelo. 2002.

KIM, Yohan, et al. Strategies for imputation of high-resolution environmental data in clinical randomized controlled trials. International Journal of Environmental Research and Public Health, 2022, vol. 19, no 3, p. 1307.

ORTH, Rene, et al. High-resolution European daily soil moisture derived with machine learning (2003–2020). Scientific Data, 2022, vol. 9, no 1, p. 1-13. doi: 10.1038/s41597-022-01785-6. PMID: 36376361; PMCID: PMC9663700

PALOMINOS-RIZZO, Teresa, et al. Estimación de la humedad del suelo mediante regresiones lineales múltiples en Llano Brenes, Costa Rica. Agronomía Mesoamericana, 2022, vol. 33, no 2, p. 13.

PRAKASH, Shikha; SAHU, Sitanshu Sekhar. Soil moisture prediction using shallow neural network. International Journal of Advanced Research in Engineering and Technology, 2020, vol. 11, no 6. Disponible en: SSRN: https://ssrn.com/abstract=3656915