Inspira

Crea

Transforma

UNIVERSIDAD EAFIT

Escuela de
Ciencias Aplicadas
e Ingeniería

www.eafit.edu.co

VIGILADA | MINEDUCACIÓN

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

Análisis de la humedad del suelo usando datos meteorológicos

GRUPO 10

Manuela Ramos Ospina — Universidad EAFIT. Pasante de investigación

Camila Acosta Gómez — Yamaha. Científica de datos

John Zapata Jimenez — Bancolombia. Científico de datos

Dany Palacio Agudelo — Servicios Ambientales y Geográficos S.A. Asistente de coordinación

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Contenido

UNIVERSIDAD EAFIT

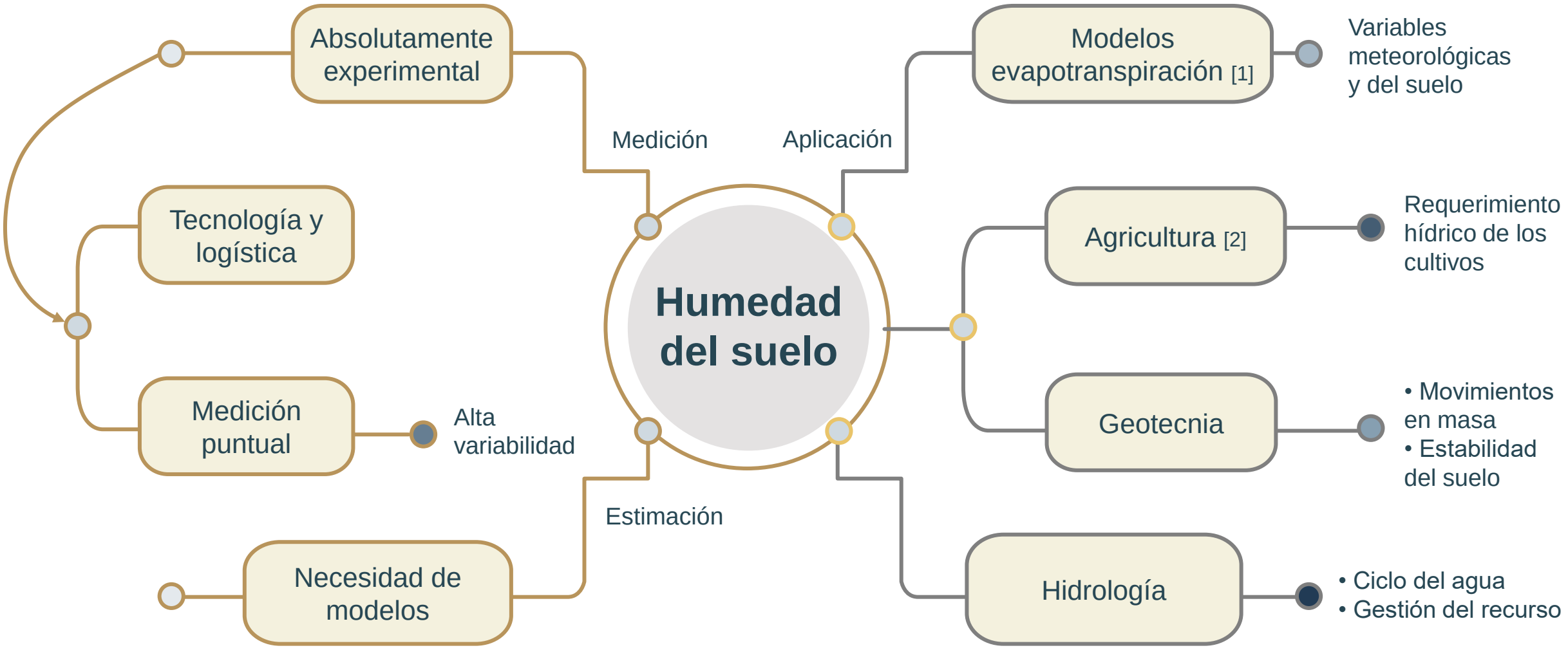
Escuela de Ciencias Aplicadas e Ingeniería

Contextualización

1

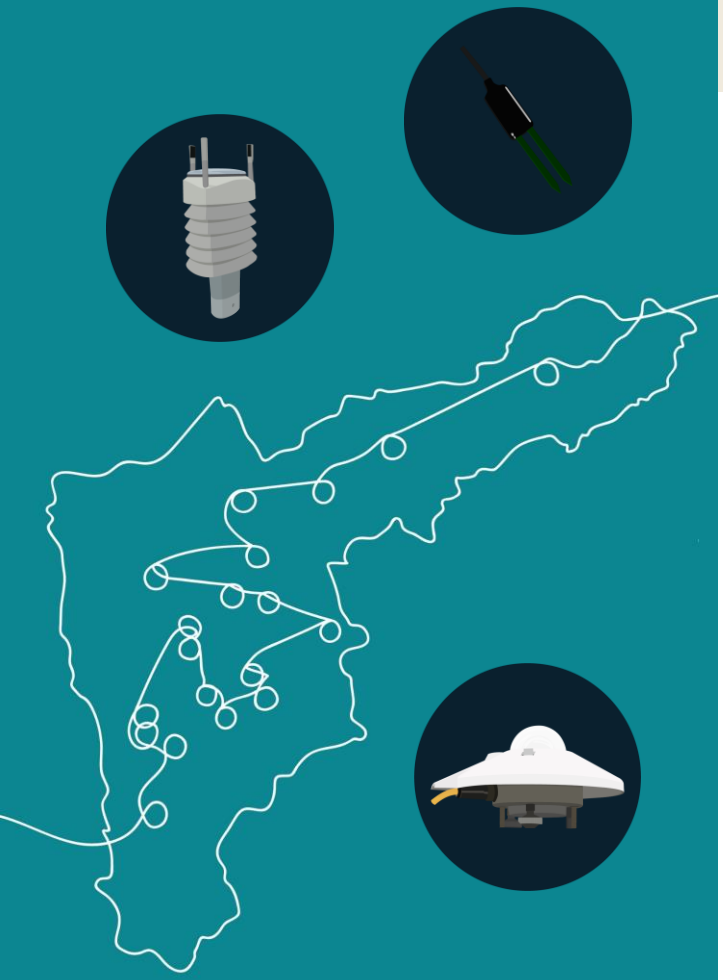
1.0 Problema de investigación

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería



[1] Ortiz, R. S., & Chile A., M. (2020). Métodos de cálculo para estimar la evapotranspiración de referencia para el Valle de Tumbaco. Siembra, 7(1), 070–079.
[2] Anello, M., Bittelli, M., Bordoni, M. et al. (2024). Robust Statistical Processing of Long-Time Data Series to Estimate Soil Water Content. Math Geosci 56, 3–26

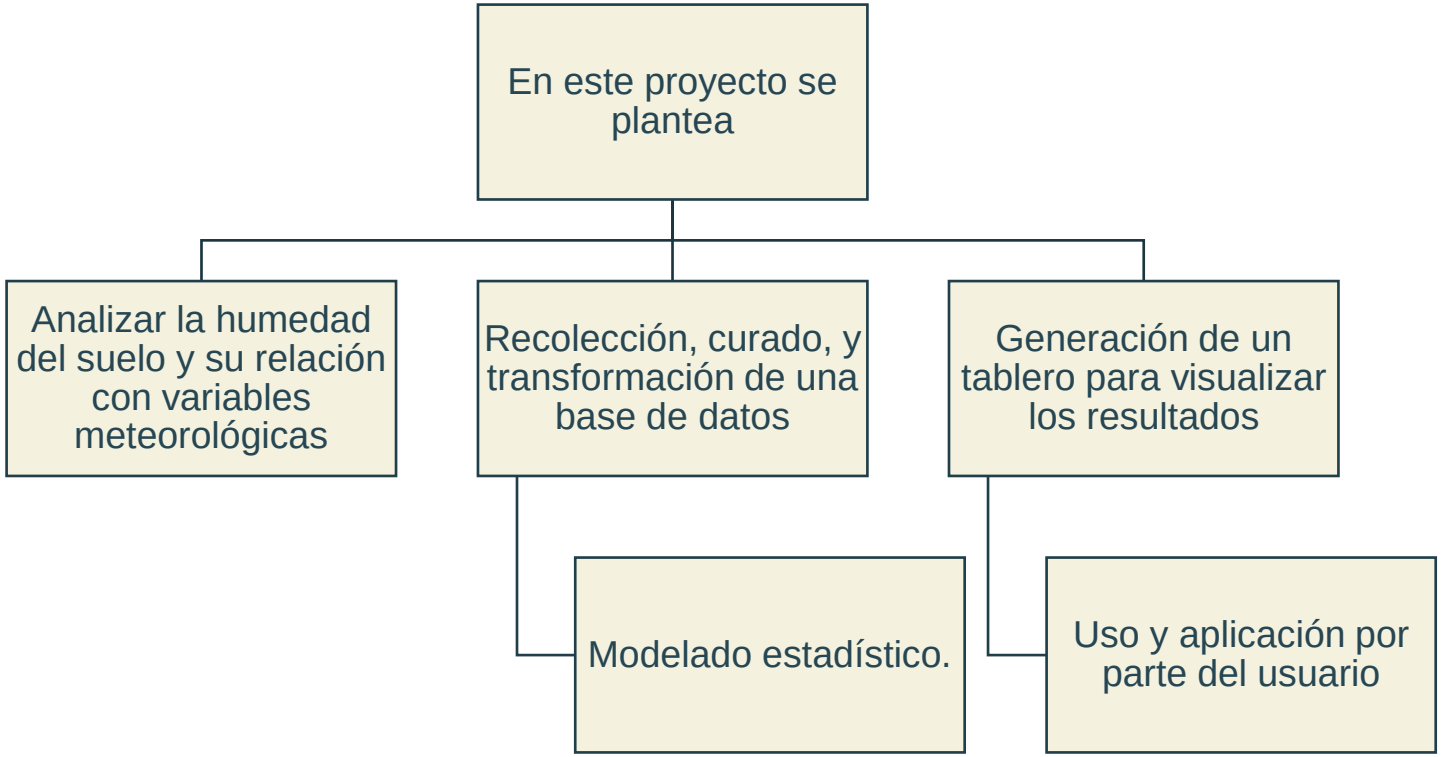
1.1 Objetivo



Redes de sensores del SIATA

Imagen adaptada de la fuente original: https://siata.gov.co/sitio_web/index.php/monitoreo#meteorologicas

Las redes de sensores meteorológicas proveen información constante y en muchos casos de forma pública

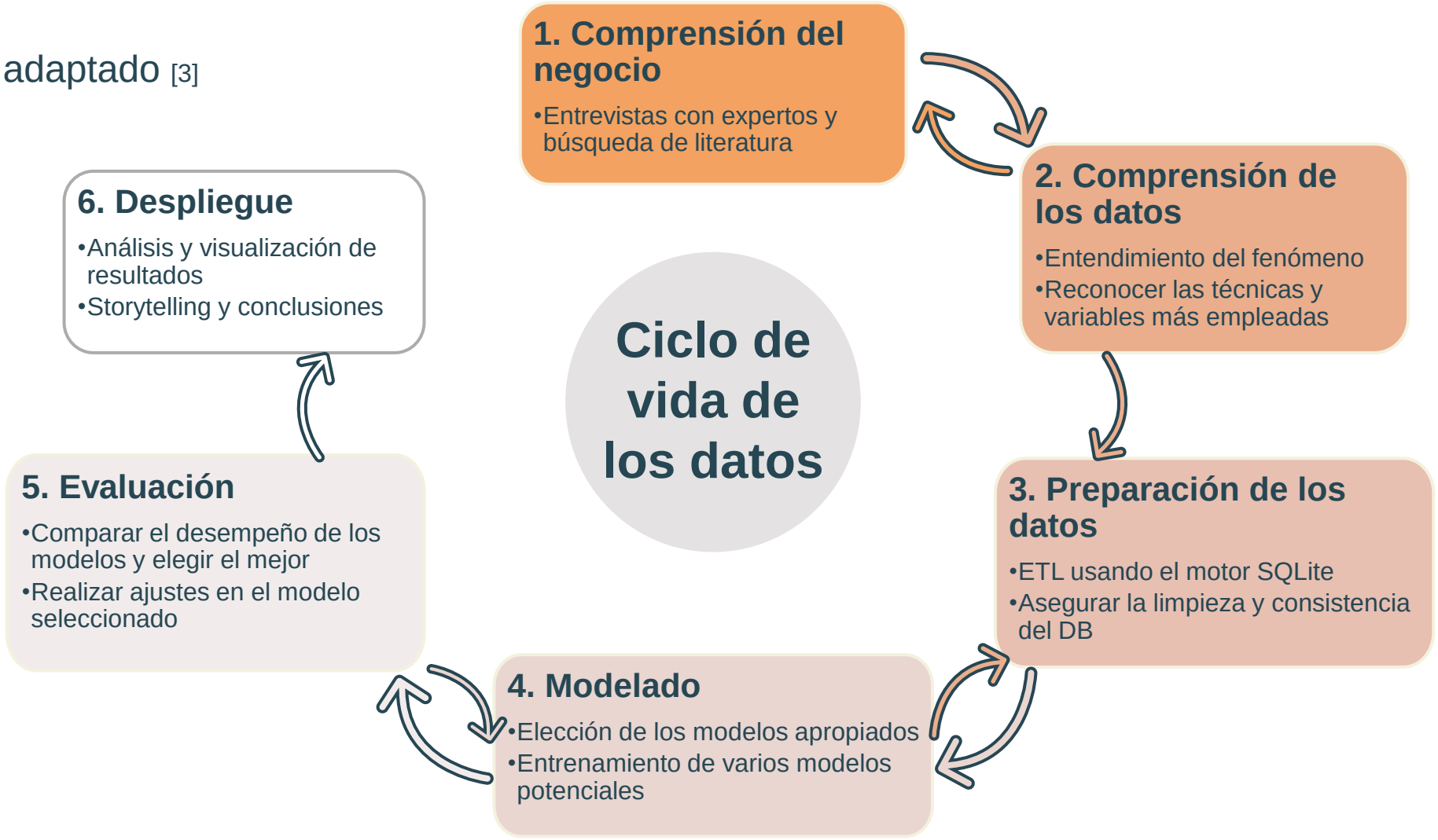


1.2 Metodología

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

CRISP-DM adaptado [3]



[3] 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/

1.3 Fuentes de datos

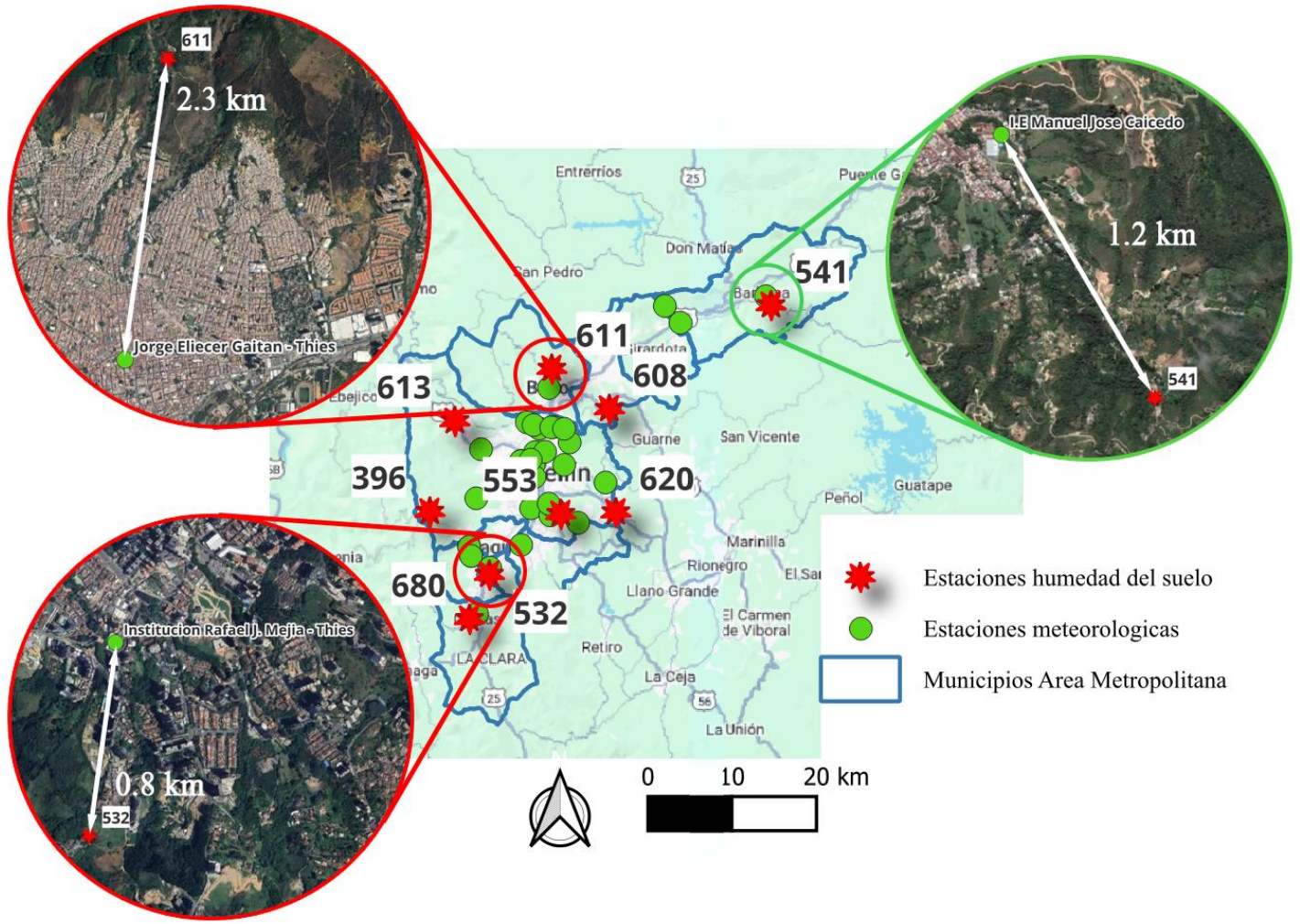
UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

IDEAM

WorldClim
Global climate and weather data

x Fata de datos
x Complejidad fuera del alcance

SIATA
Área Metropolitana Valle de Aburrá
SISTEMA DE ALERTA TEMPRANA



✓ Acceso es público
✓ Cubrimiento del área metropolitana
✓ Histórico de los datos

cómo se realizó la medida de distancia?

Contenidos

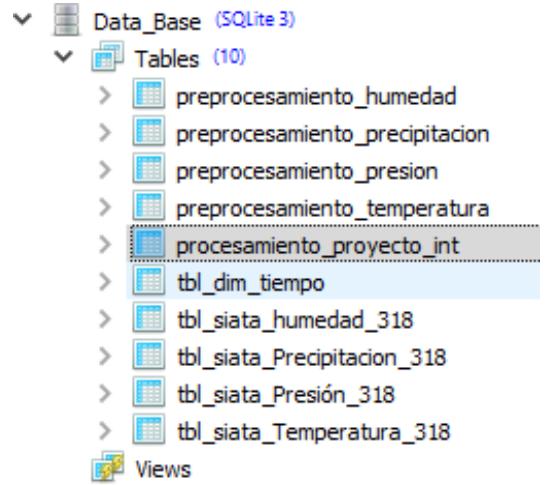
UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Desarrollo metodológico

2

2.0 Recuperación de los datos

ETL	• Ingesta en batch • Conexión mediante SQLite
Ingeniería de datos	• Tabla con dimensión temporal • Construcción variables exógenas
Preparación de los datos	• Limpieza índices de calidad dudosa
Despliegue y creación del DB	• Query integrador



[4] 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

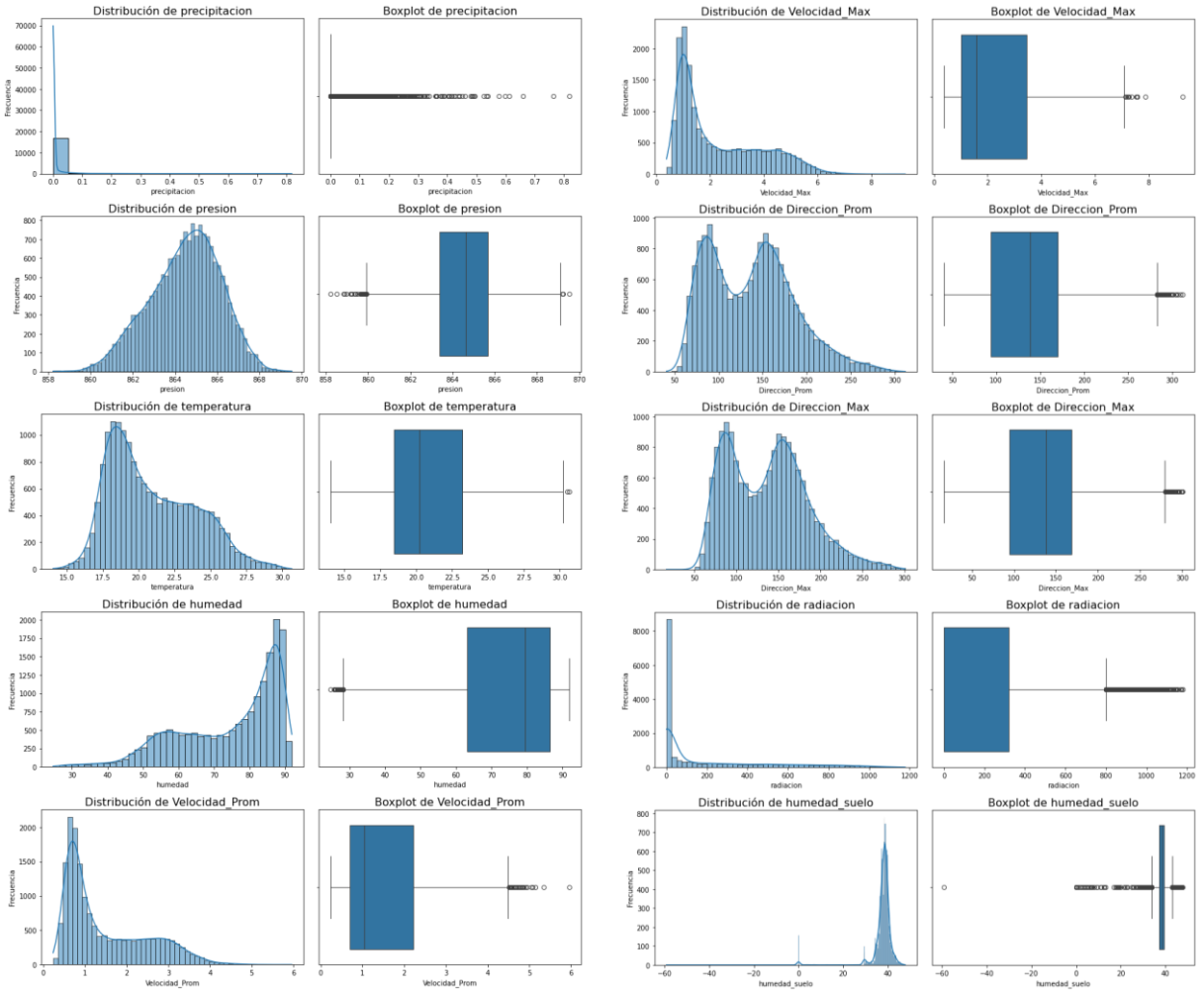
UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

2.1 Análisis exploratorio

Índice	Nombre de la variable	Unidades
1	Precipitación	mm
2	Presión atmosférica	hPa
3	Temperatura	°C
4	Humedad relativa	%
5	Magnitud de la velocidad promedio del viento	m/s
6	Magnitud de la Velocidad Máxima del viento	m/s
7	Dirección promedio del viento	grados
8	Dirección Máxima del viento	grados
9	Radiación solar	W/m2
10	Contenido de humedad del suelo	m3/ m3
11	Fecha y hora	A-m-d H:M:S
12	Mes	m
13	Semana año	semana

Rango de fechas: sep 2021 – sep 2023

Gráficos de distribución y boxplot para cada variable ambiental [4]

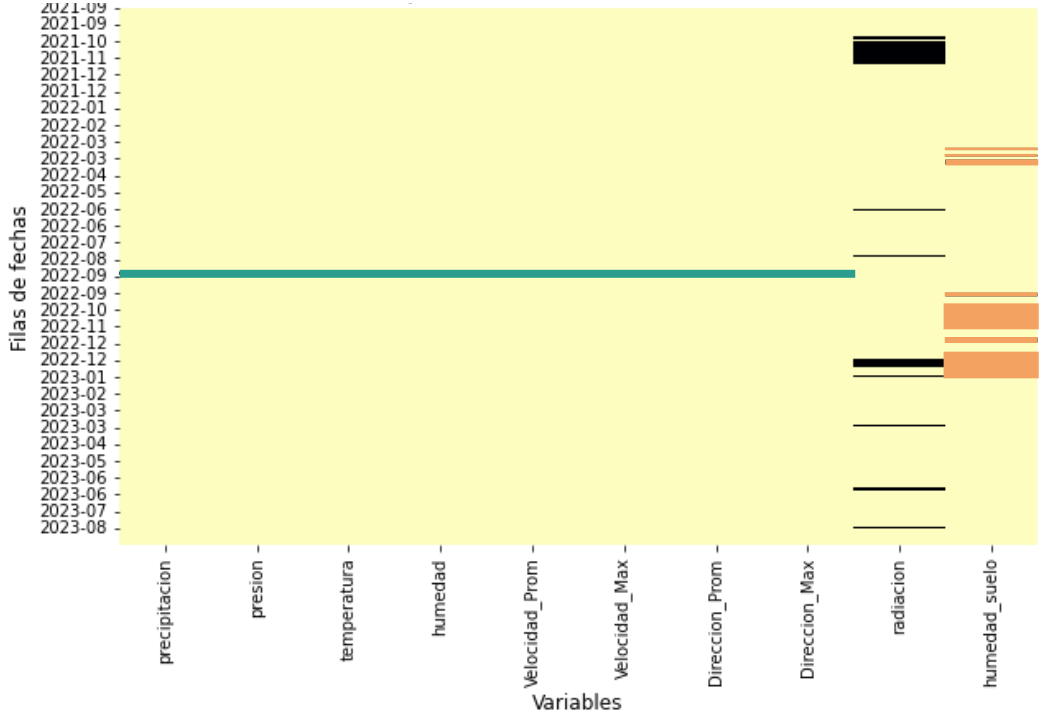


2.1.1 Preparación de los datos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Identificación y rellenado de datos nulos

Mapa de calor para los datos nulos y no nulos



Índice	Nombre de la variable	Datos nulos
1	Precipitación	100
2	Presión atmosférica	100
3	Temperatura	100
4	Humedad relativa	100
5	Magnitud de la velocidad promedio del viento	100
6	Magnitud de la Velocidad Máxima del viento	100
7	Dirección promedio del viento	100
8	Dirección Máxima del viento	100
9	Radiación solar	1,422
10	Contenido de humedad del suelo	2,130

Baseline: Median Imputation (MI)

Multivariado: Multiple Imputation by Chained Equations (MICE) [5]

Regresión: Linear regression

[5] 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.

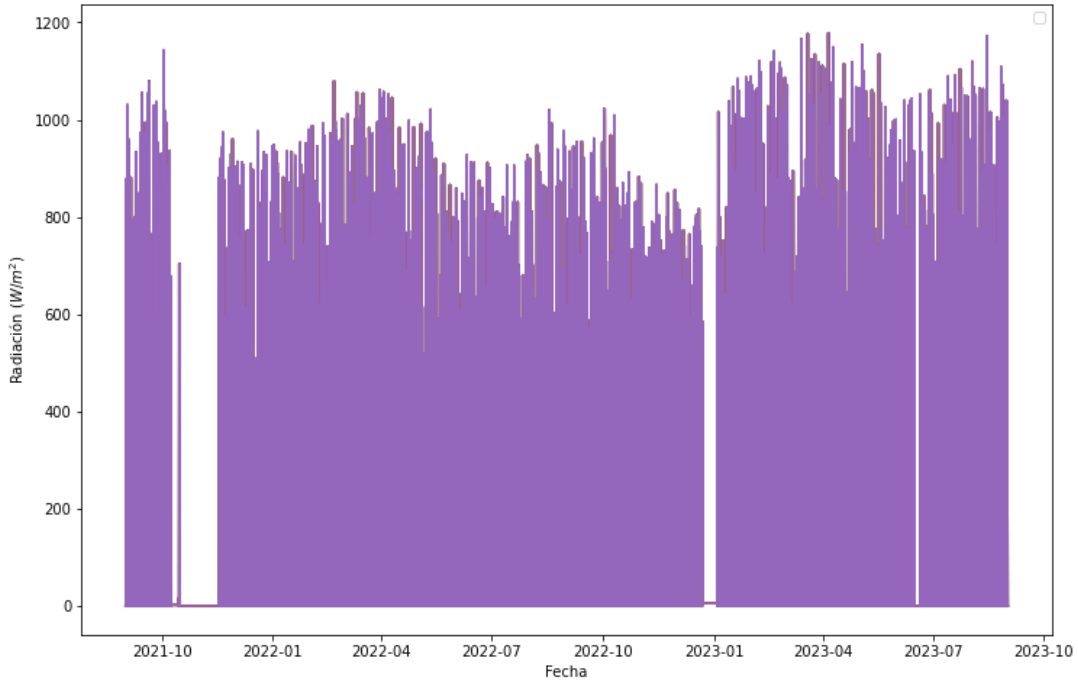
2.1.1 Preparación de los datos

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

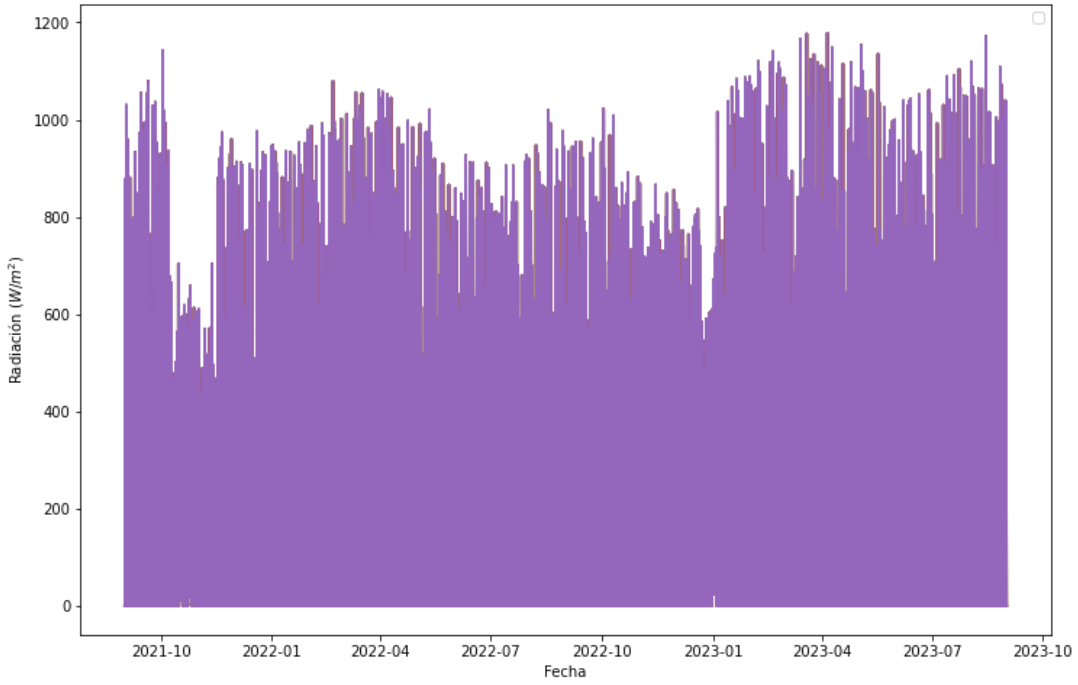
Identificación y rellenado de datos nulos

Serie temporal de radiación explorando dos técnicas de inputado

inputación con MI



inputación con MICE



2.1.1 Preparación de los datos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Identificación y rellenado de datos nulos

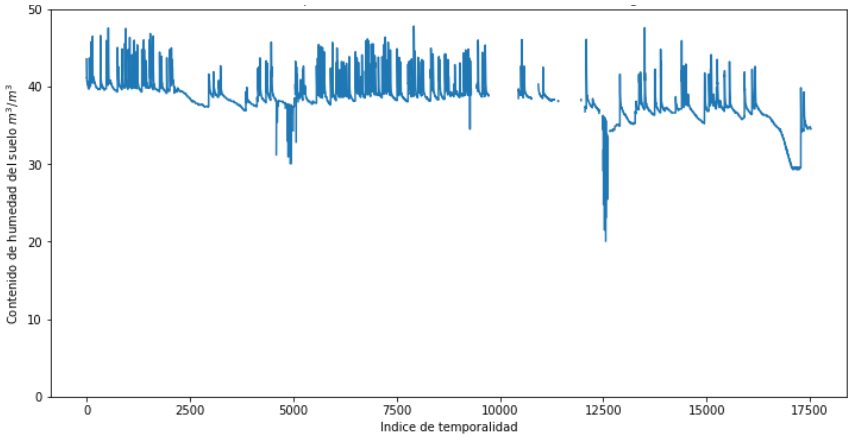
Inputación de datos de humedad del suelo con regresión lineal

Matriz de correlación

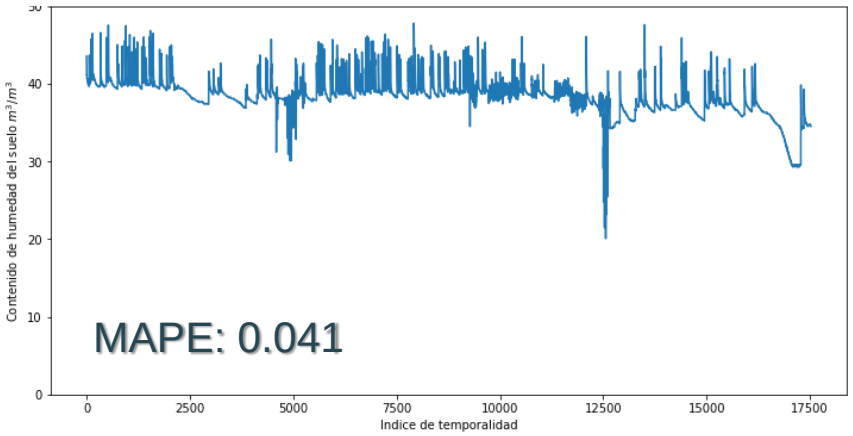
	semana_año	precipitacion	presion	temperatura	humedad	Velocidad_Prom	Velocidad_Max	Direccion_Prom	Direccion_Max	radiacion	humedad_suelo
semana_año	1	0.0077	0.0069	0.0069	0.03	-0.011	-0.014	0.025	0.022	-0.038	0.24
precipitacion	0.0077	1	0.039	-0.1	0.13	0.0072	0.0048	0.075	0.074	-0.094	0.077
presion	0.0069	0.039	1	-0.47	0.39	-0.5	-0.49	0.34	0.34	-0.038	0.085
temperatura	0.0069	-0.1	-0.47	1	-0.93	0.75	0.78	-0.49	-0.49	0.65	-0.092
humedad	0.03	0.13	0.39	-0.93	1	-0.72	-0.75	0.43	0.43	-0.74	0.12
Velocidad_Prom	-0.011	0.0072	-0.5	0.75	-0.72	1	0.99	-0.62	-0.62	0.44	-0.045
Velocidad_Max	-0.014	0.0048	-0.49	0.78	-0.75	0.99	1	-0.59	-0.59	0.48	-0.04
Direccion_Prom	0.025	0.075	0.34	-0.49	0.43	-0.62	-0.59	1	0.99	-0.26	0.07
Direccion_Max	0.022	0.074	0.34	-0.49	0.43	-0.62	-0.59	0.99	1	-0.25	0.069
radiacion	-0.038	-0.094	-0.038	0.65	-0.74	0.44	0.48	-0.26	-0.25	1	-0.044
humedad_suelo	0.24	0.077	0.085	-0.092	0.12	-0.045	-0.04	0.07	0.069	-0.044	1

1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

Serie temporal de humedad del suelo original



Serie temporal de humedad del suelo inputado



2.1.1 Preparación de los datos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Transformación de las variables

- Corrección de asimetría (skewness)
- Corrección de sesgos
- Corrección de no-linealidad

Verificar positividad → **Estimar λ** → **Aplicar transformación** → **Finalización**

$x > 0 \Rightarrow$ Box-Cox

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

Optimiza la función de verosimilitud para λ

- Datos transformados
- Valor óptimo de λ
- valor la asimetría

$x \leq 0 \Rightarrow$ Yeo-Johnson

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

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

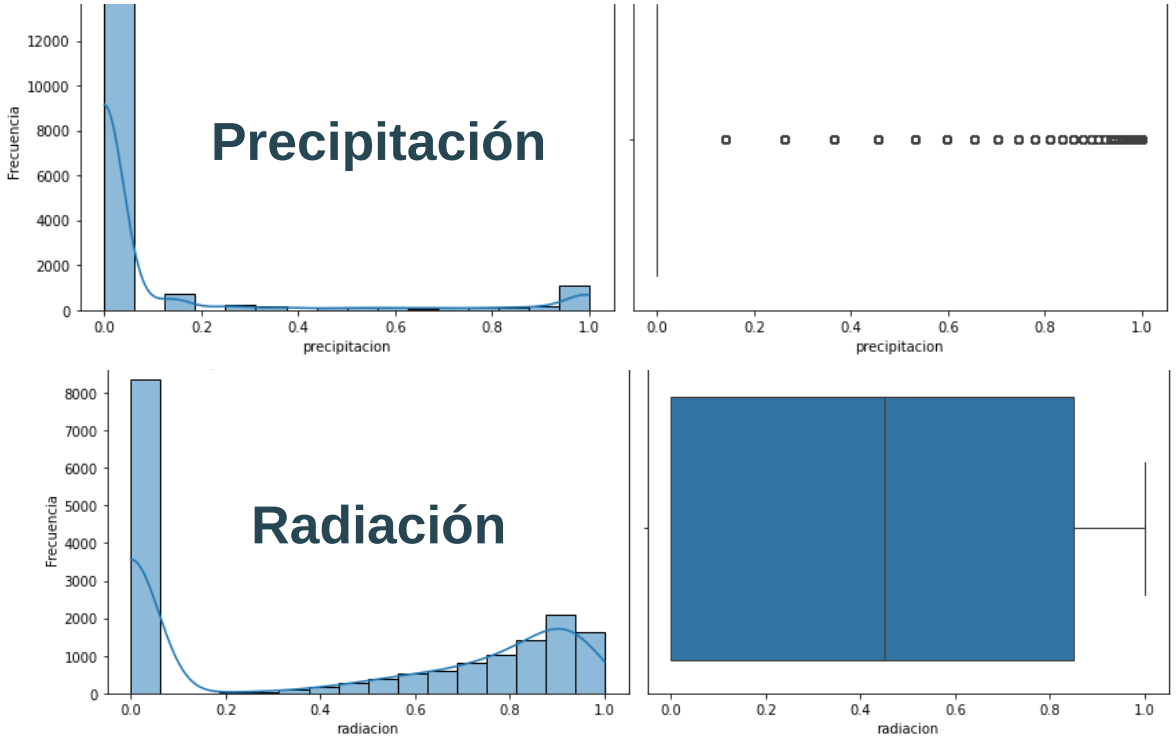
2.1.1 Preparación de los datos

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Estandarización de las variables

Escalamiento min-max: $$0 \leq \frac{X - X_{min}}{X_{max} - X_{min}} \leq 1$$

Distribución variables precipitación y radiación

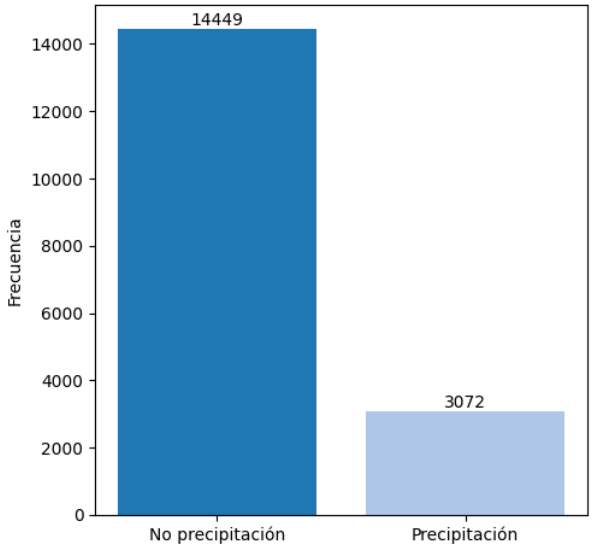


Balanceo de las variables

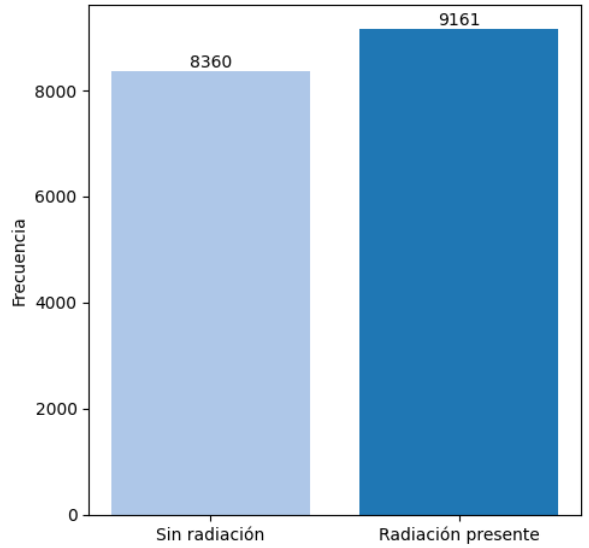
- Reemplazo de outliers de humedad del suelo
- Eliminación de variables redundantes
- Binarización de precipitación y radiación

Distribución de etiquetas binarizadas

Precipitación binarizada



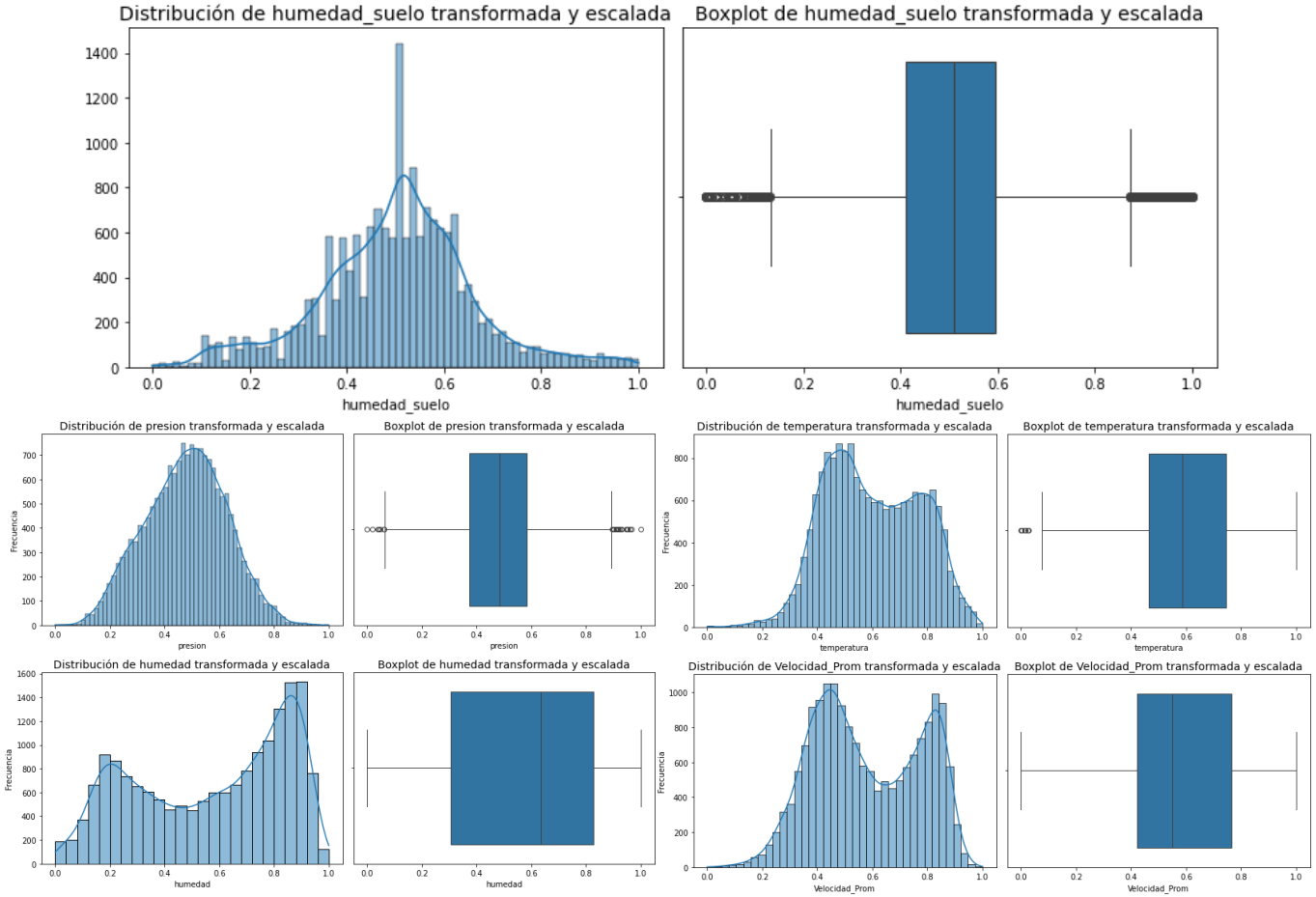
Radiación binarizada



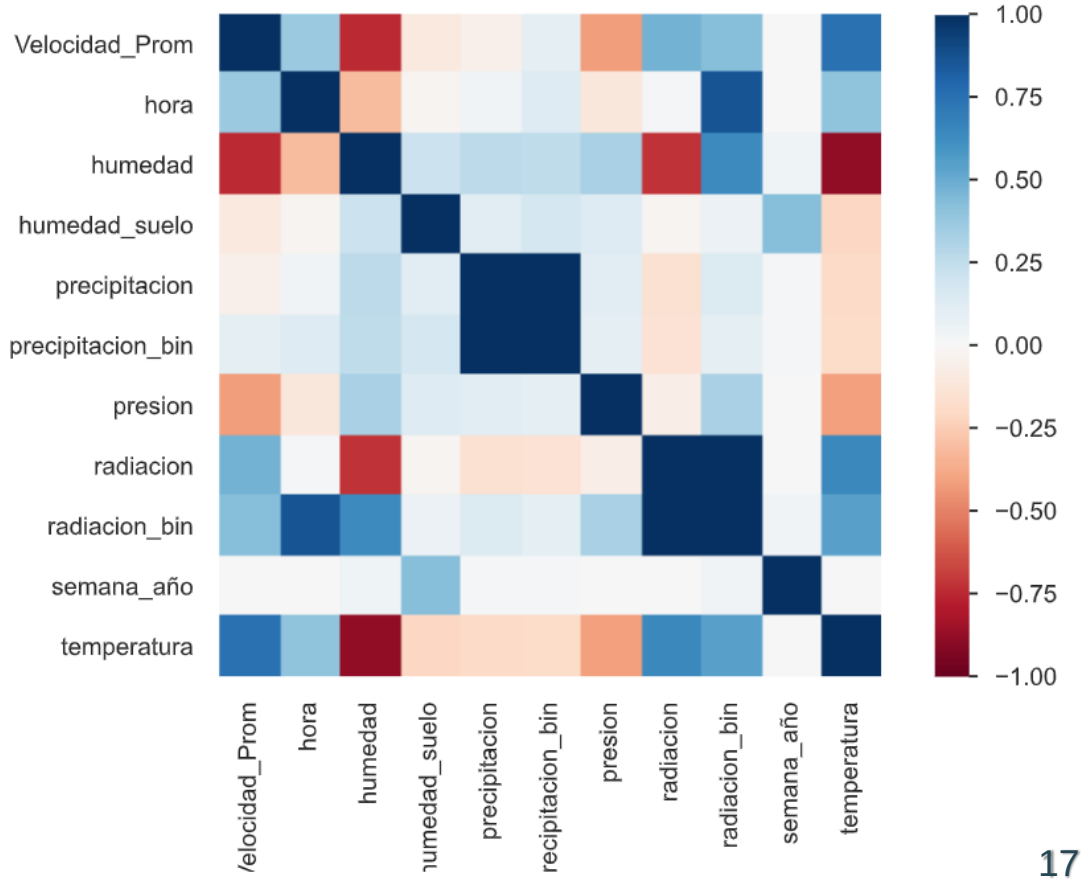
2.1.2 Análisis descriptivo de los datos

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería

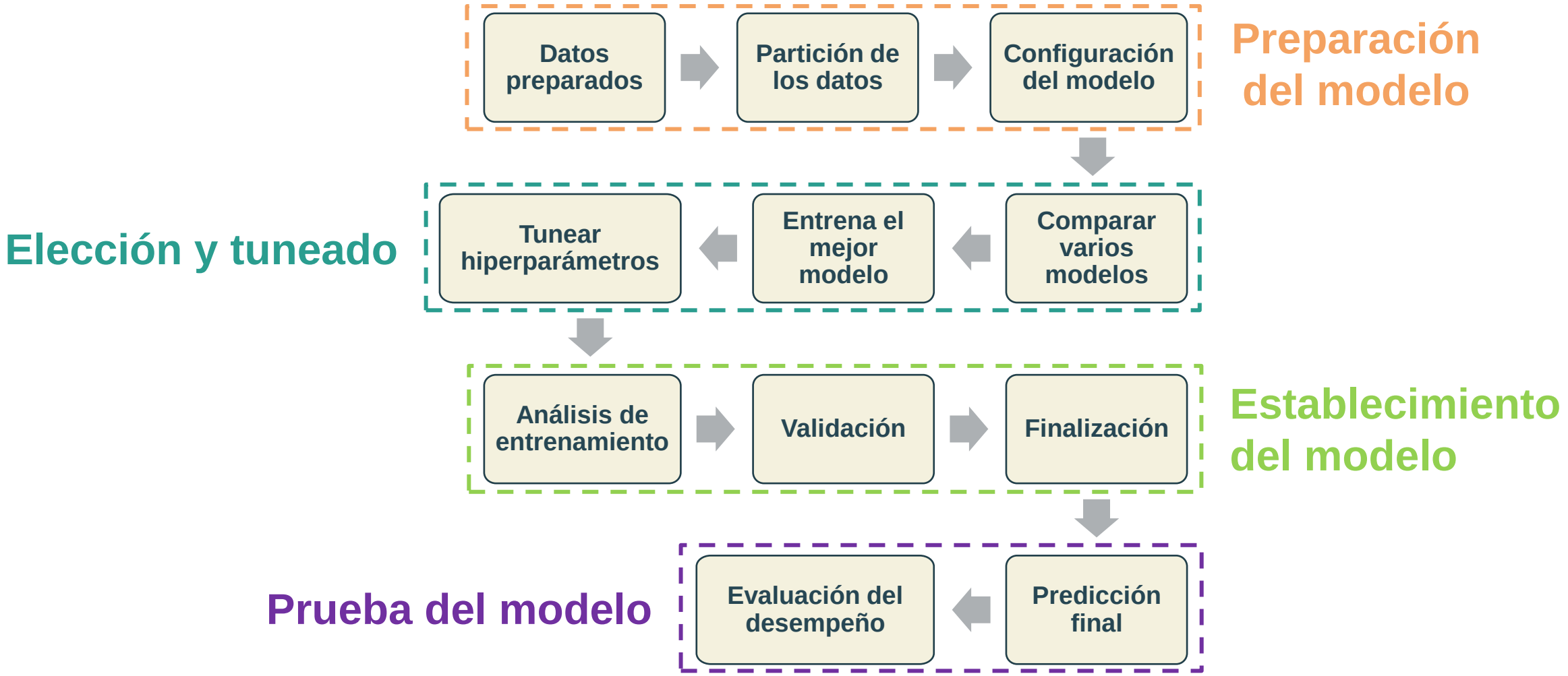
Gráficos de distribución y boxplot para las variables transformadas y escaladas



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



2.2 Modelación de los datos



2.2.1 Preparación del modelo

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Transformación & Escalado

Datos preparados → **Partición de los datos** → **Configuración del modelo**

Conjunto de datos
Variables ambientales + hora como variable categórica
Variables ambientales SIN hora

nov - 2021

Conjunto de prueba

- ✓ Variable respuesta
- ✓ Variable categórica
- ✓ Variables a ignorar

NOT nov - 2021

Conjunto de entrenamiento

- 70% Train
  - X train
  - Y train
- 30% Val
  - X val
  - Y val

2.2.2 Elección y tuneado

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Comparar varios modelos → **Entrena el mejor modelo** → **Tunear hiperparámetros**

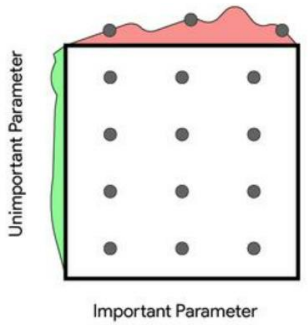
- ✓ Entrenamiento de varios modelos a la vez
- ✓ Evaluar con métricas de desempeño (R2, MAPE, TT (Sec))
- ✓ Identificar el mejor modelo

Modelos comparados
KNN regresor
Regresión Lineal
Regresión Ridge
Regresión Lasso
Elastic Net

- ✓ Entrenamiento de los mejores modelos
- ✓ Validación cruzada 8 fold

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

- ✓ Regresión lineal: intercepto y pendiente (β_0, β_i)
- ✓ KNN: K, Distancia
- ✓ *GridSearch* [6]



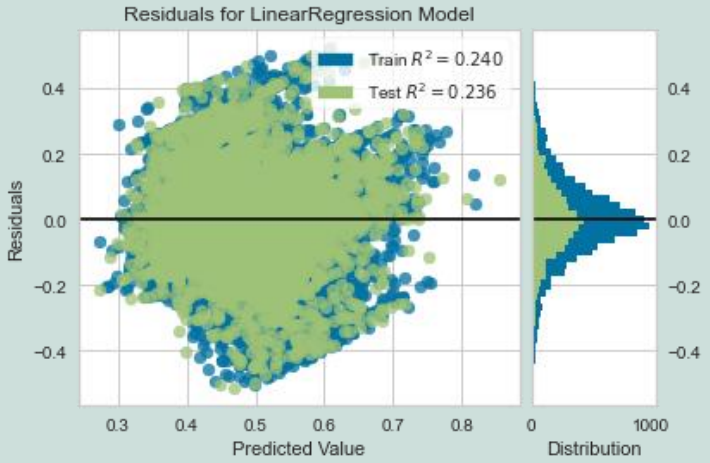
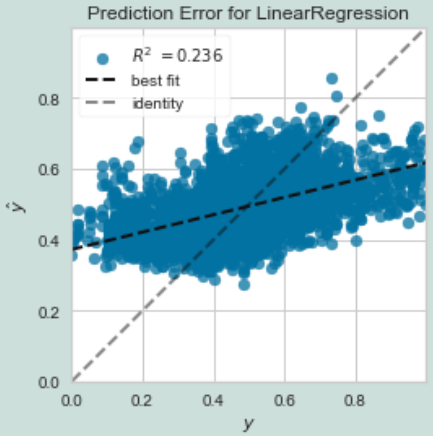
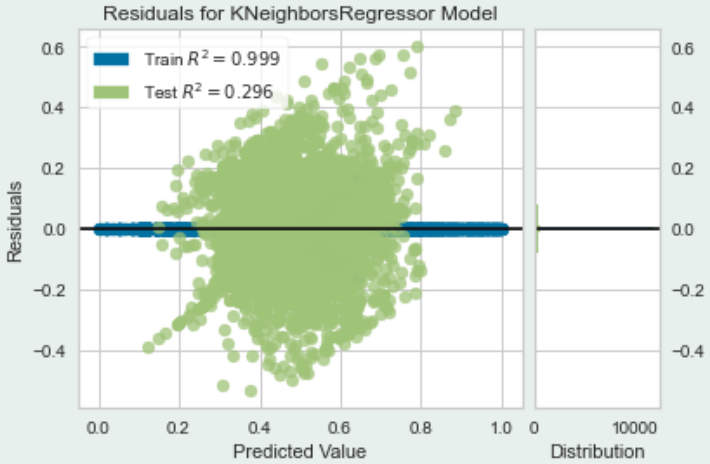
[6] Hyperparameters Optimization methods – ML. Geeks for geeks. 2023. Accedido: jun. 1, 2024 [En línea]. Disponible en: https://www.geeksforgeeks.org/hyperparameters-optimization-methods-ml/

2.2.3 Establecimiento del modelo

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

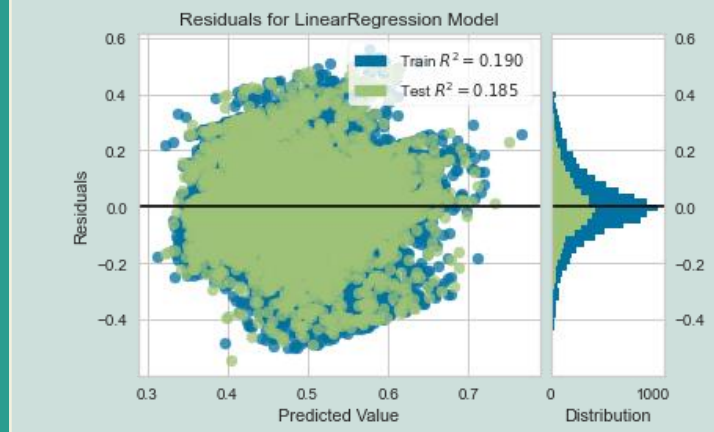
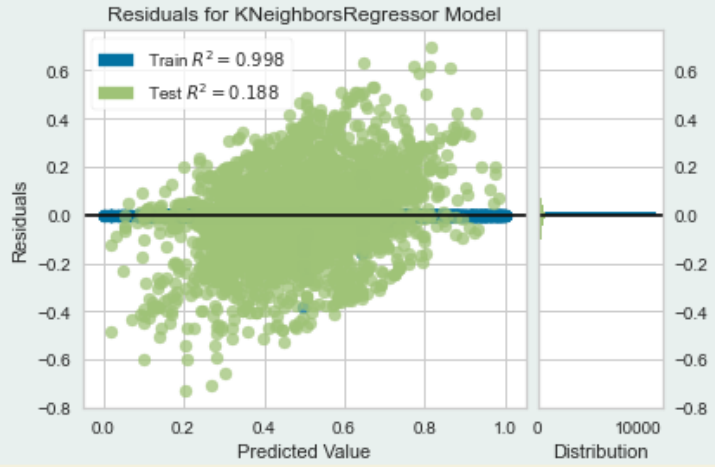
Análisis de entrenamiento

	Análisis de residuales	Error de predicción
RL1		
KNN1		

2.2.3 Establecimiento del modelo

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Análisis de entrenamiento → **Validación**

	Análisis de residuales	Error de predicción
RL2		
KNN2		

Predecir

30% Val

Finalización

Ajuste

Train + Val

Congelar hiperparámetros

2.2.4 Prueba del modelo

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

Predicción final → **Evaluación del desempeño**

nov - 2021

Conjunto de prueba

nmetric='manhattan',
n_jobs=-1,

n_neighbors=6,
weights='distance-

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		

Contenidos

UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Despliegue

3

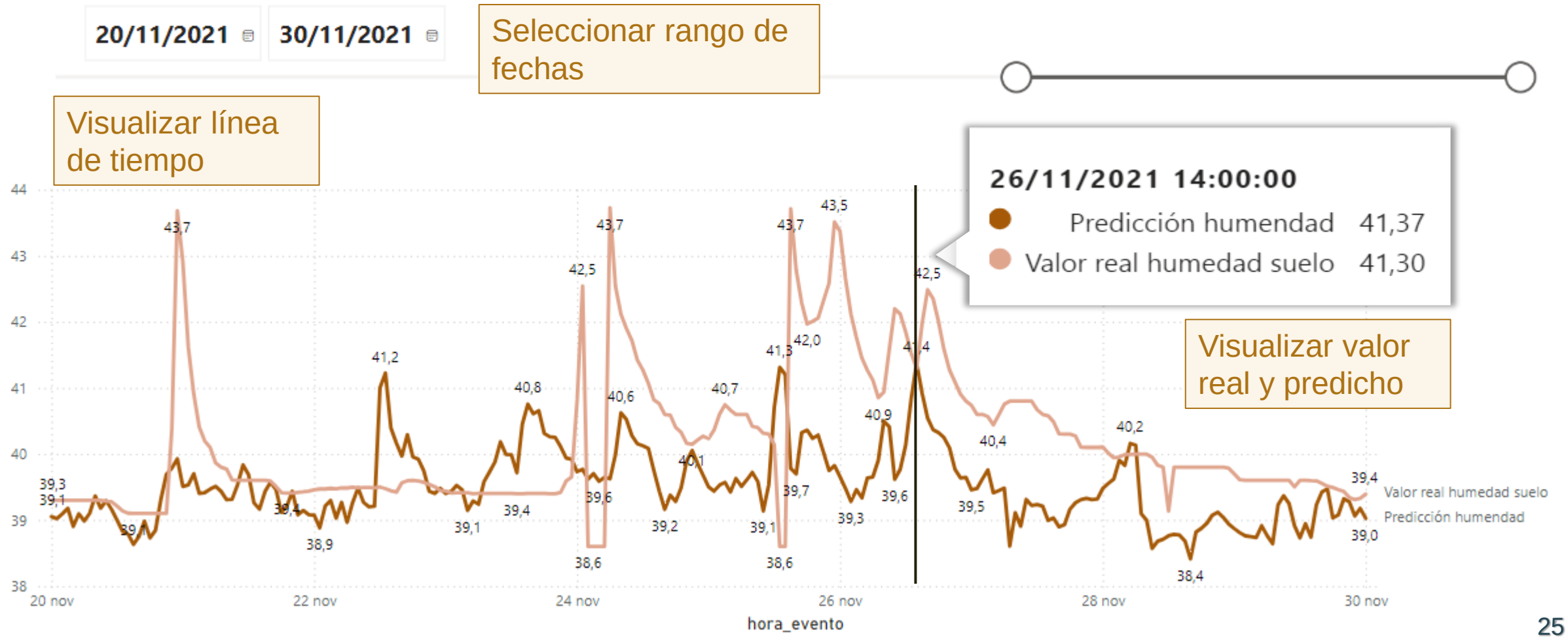
3.0 Dashboard

Power BI

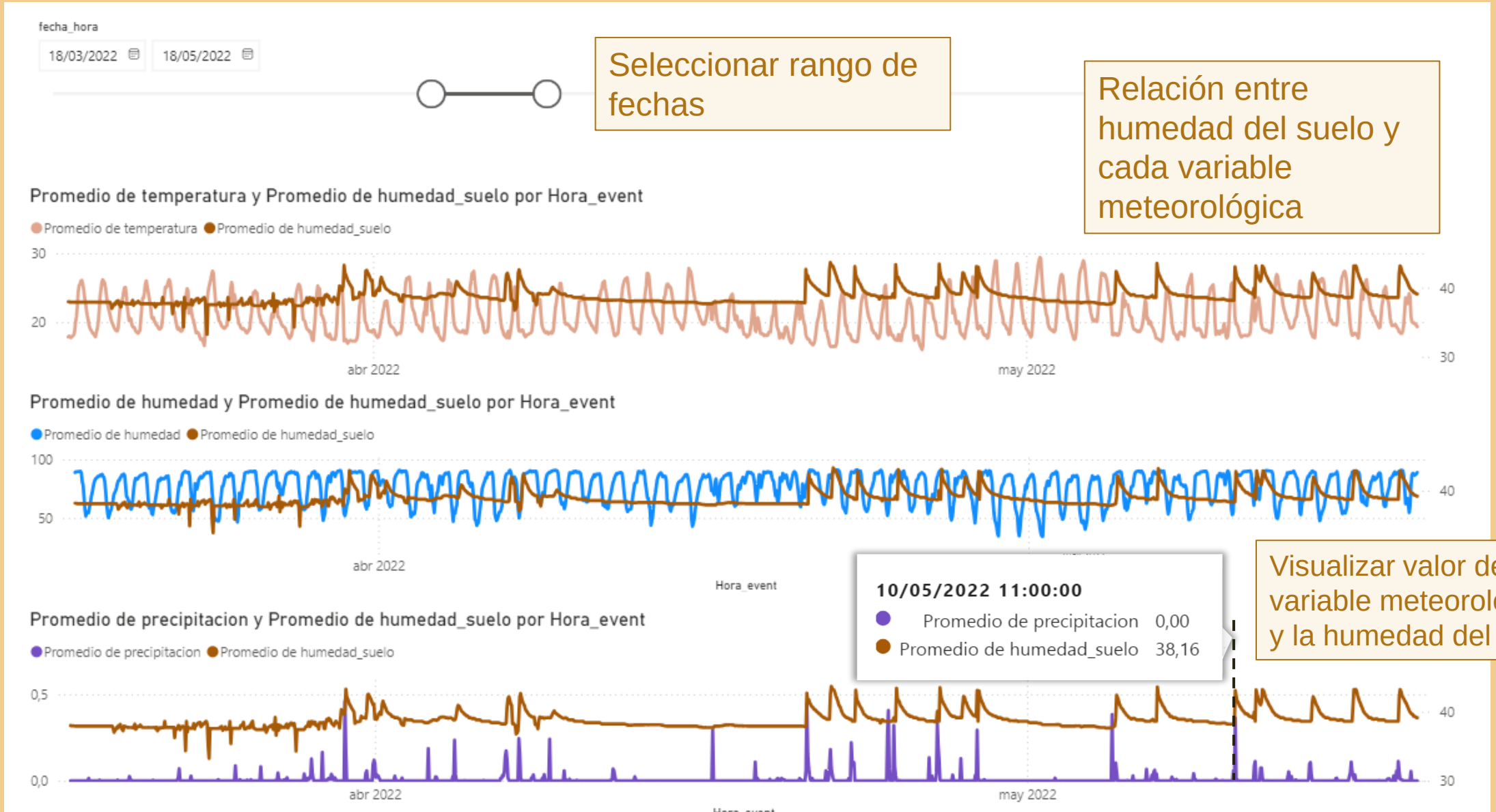
UNIVERSIDAD EAFIT

Escuela de Ciencias Aplicadas e Ingeniería

Análisis de la humedad del suelo usando datos meteorológicos



3.0 Dashboard



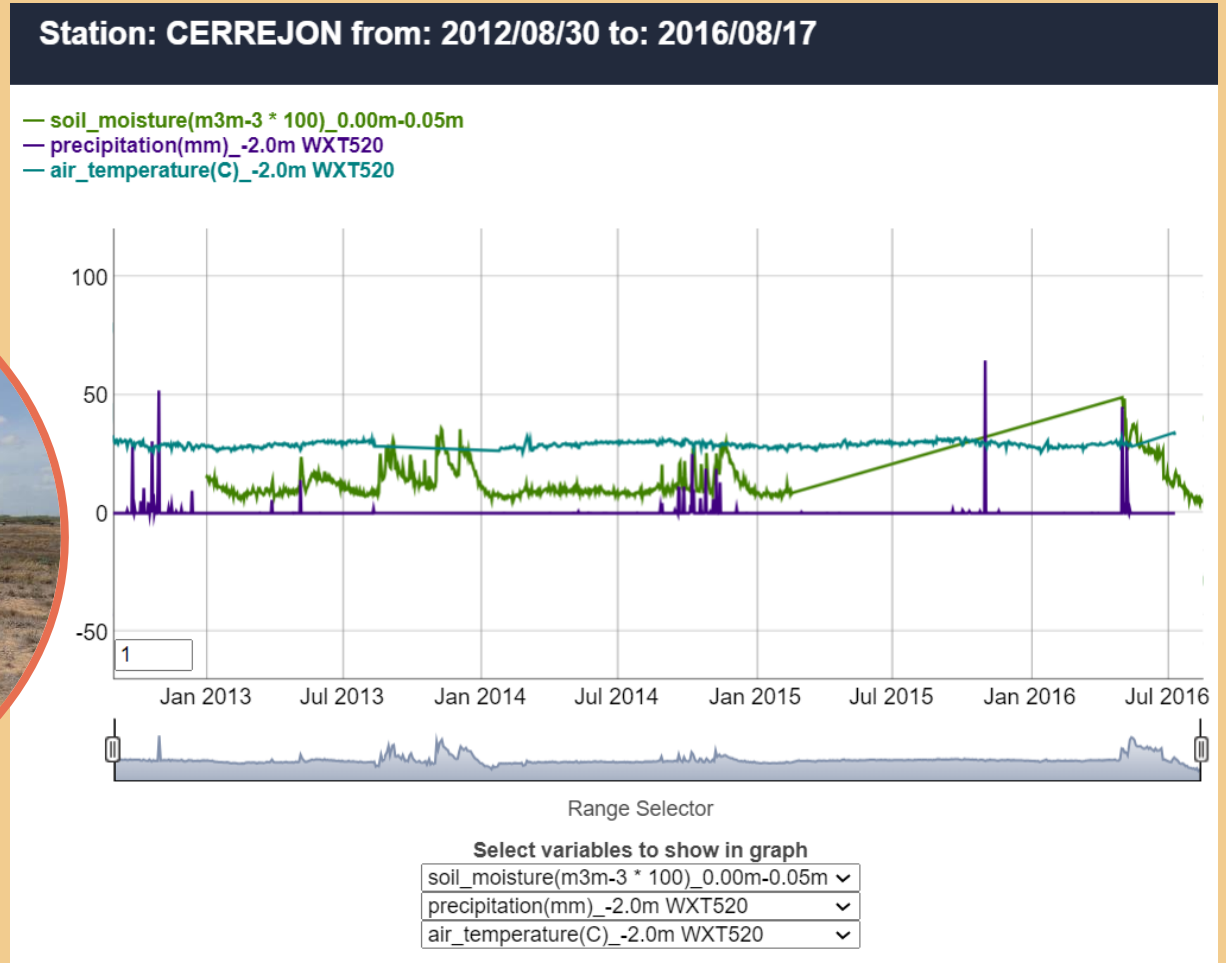
3.1 Posible caso de uso

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería

https://ismn.earth/en/dataviewer/#

3.1 Posible caso de uso

UNIVERSIDAD EAFIT | Escuela de Ciencias Aplicadas e Ingeniería



Conclusiones

4

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 R2 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.

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 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

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

La medición de la humedad del suelo es un tema aún experimental, 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

¡Gracias!