EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 152 No. 6
June 2023

Research in Computing Science

Research in Computing Science, Año 22, Volumen 152, No. 6, junio de 2023, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de junio de 2023.

Research in Computing Science, year 22, Volume 152, No. 6, June 2023, is published monthly by the Center for Computing Research of IPN.

Volume 152(6)

Advances in Artificial Intelligence

Gilberto Ochoa Ruiz (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2023

ISSN: in process

Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Aplicaciones de aprendizaje automático para estimar la evaporación en regiones áridas: caso de estudio Calera, Zacatecas

Luis Fernando Castillo Martínez, Julián González Trinidad, Hugo Enrique Júnez Ferreira, Carlos Francisco Bautista Capetillo, Cruz Octavio Robles Rovelo, José Armando Rodríguez Carrillo

Universidad Autónoma de Zacatecas,
Campus UAZ Siglo XXI,
Unidad Académica de Ingeniería Eléctrica,
México

{fercast, jgonza, hugo.junez, baucap, octavio.robles, jarmando.rc}@uaz.edu.mx

Resumen. La evaporación es un proceso fundamental dentro del ciclo hidrológico, el cual consiste en la pérdida de agua en forma de vapor desde la superficie terrestre hacia la atmósfera. Debido a su complejidad, se han implementado diferentes técnicas de Aprendizaje Automático (ML, por sus siglas en inglés), para comprender mejor este proceso. En esta investigación se realizó una comparación de tres modelos de ML, regresión lineal múltiple (MLR), bosques aleatorios (RF) y k-vecinos más cercanos (KNN), para estimar la evaporación en la región Calera, Zacatecas, México. Para evaluar el rendimiento de los modelos, se utilizaron las métricas coeficiente de correlación de Pearson (R), coeficiente de eficiencia Nash-Sutcliffe (NSE), raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE). El modelo regresión lineal múltiple (MLR) fue el que presentó mejor desempeño, con un coeficiente de correlación de Pearson (R) para la estación Calera de 0.97 y para Fresnillo de 0.94, de igual manera, se obtuvo un NSE de 0.93 y 0.87, un RMSE de 15.97 y 20.53 mm, y un MAE de 12.56 y 14.66 mm, respectivamente.

Palabras clave: Evaporación, aprendizaje automático, regresión lineal múltiple, bosques aleatorios, k-vecinos más cercanos.

Machine Learning Applications for Evaporation Estimation in Arid Regions: A Case Study in Calera, Zacatecas

Abstract. Evaporation is a key process in the hydrological cycle, consisting of the loss of water in vapor form from the land surface to the atmosphere. Due to its complexity, various Machine Learning (ML) techniques have been developed to better understand this phenomenon. In this research, it was compared the performance of three ML models, multiple linear regression (MLR), random

forest (RF), and k-nearest neighbors (KNN), for estimating evaporation in the region of Calera, Zacatecas, Mexico. To evaluate model performance, it was used the Pearson correlation coefficient (R), Nash-Sutcliffe efficiency coefficient (NSE), root mean square error (RMSE), and mean absolute error (MAE). The results showed that multiple linear regression performed best in the study area, with a Pearson correlation coefficient (R) of 0.97 for the Calera climatological station and 0.94 for Fresnillo. The NSE values were 0.93 and 0.87, the RMSE values were 15.97 and 20.53 mm, and the MAE values were 12.56 and 14.66 mm, respectively.

Keywords: Evaporation, machine learning, multiple linear regression, random forest, k-nearest neighbors.

1. Introducción

La evaporación es uno de los componentes más importantes del ciclo hidrológico, en el cual, el agua en su fase líquida parte de la superficie de la tierra hacia la atmósfera en forma de vapor de agua [2], es probablemente el parámetro más complicado y difícil de estimar de entre todos los elementos que integran el ciclo hidrológico debido a las interacciones complejas de los componentes hidrológicos como la superficie del agua, el suelo, el proceso atmosférico y la vegetación.

Por lo tanto, le estimación de la evaporación es un tema relevante en el manejo de los recursos hídricos y la agricultura, particularmente en regiones áridas y semi-áridas. Debido a la diferencia de temperatura, este fenómeno hidrológico es un proceso no lineal que ocurre en la naturaleza [13].

Este fenómeno es influenciado por el suministro de energía calórica y el gradiente de vapor de presión, los cuales están relacionados con datos meteorológicos tales como la temperatura del aire, radiación solar, humedad relativa, velocidad del viento y la presión atmosférica, a su vez estos aspectos están estrechamente relacionados con otros factores como la ubicación geográfica, la hora del día, la temporada del año y el tipo de clima [1].

La estimación de la evaporación en áreas áridas y semi-áridas es de suma importancia debido a la poca disponibilidad de agua en las fuentes de abastecimiento, así como, para los requerimientos de agua de la vegetación en los ecosistemas [16].

La pérdida de agua por evaporación se ha incrementado significativamente durante las últimas décadas, particularmente en regiones áridas y semi-áridas a lo largo del mundo. Por lo tanto, la estimación precisa de las tasas de evaporación es vital para diferentes contextos, como el presupuesto y manejo del agua para irrigación, hidrología, agronomía y manejo de los recursos hídricos.

Generalmente, la evaporación del agua superficial es medida utilizando dos métodos, el primero a través de una medición directa mediante tanques evaporímetros, y la segunda utilizando ecuaciones semi-empíricas basadas en variables climáticas generando una medición indirecta [10]. Existe una gran variedad de tanques evaporímetros que tienen diferente forma y tamaño, sin embargo, el tanque evaporímetro estándar clase A es uno de los tanques utilizados más comunes y con una aceptación a nivel global [8].

Tabla 1. Coordenadas de las estaciones climatológicas.

Estación	Latitud (Norte)	Longitud (Oeste)	Municipio
El Pardillo 3	$22°54'31''$	$102°39'34''$	Fresnillo
CEZAC	$22°10'49''$	$102°43'1''$	Calera
Fresnillo	$23°10'40''$	$102°53'20''$	Fresnillo
Calera	$22°54'00''$	$102°39'00''$	Calera

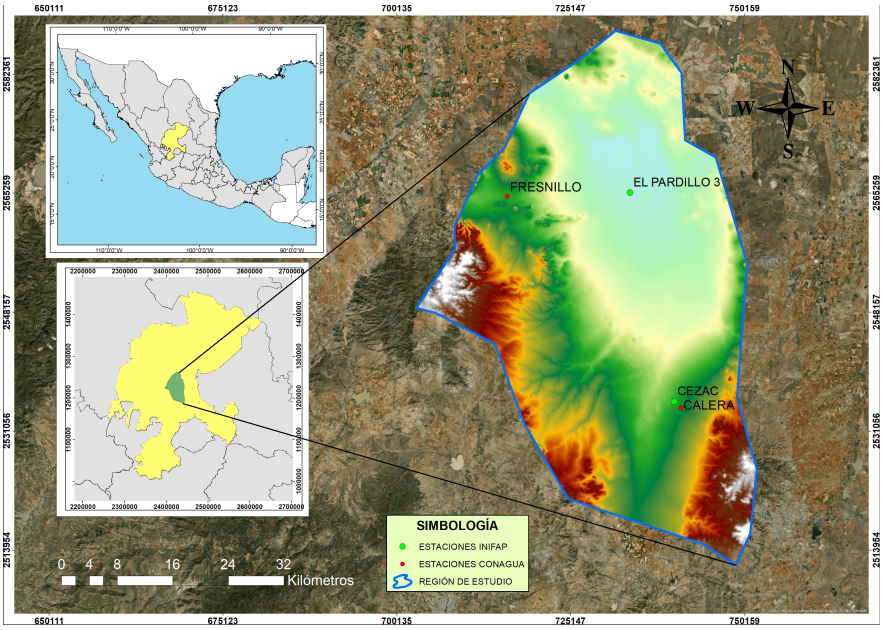

Fig. 1. Estaciones climatológicas de la región Calera.

El desarrollo de métodos de estimación indirecta basados en el uso de diferentes variables meteorológicas tales como las horas de luz solar, velocidad del viento, humedad relativa, precipitación, temperatura máxima, mínima y media a menudo son sugeridos para estimar la evaporación, especialmente cuando se trabaja con modelos empíricos y semi-empíricos.

Sin embargo, una de las limitantes para estimar la evaporación es la naturaleza dinámica de las variables meteorológicas aplicadas, debido a que es no estacionario y presenta características estocásticas. Por tanto, se requiere el desarrollo de modelos inteligentes, robustos y confiables para estimar la evaporación, el desarrollo de tales modelos se ha incrementado dentro del campo de la ingeniería y administración de recursos hídricos [18].

En los últimos años, investigadores han tratado de modelar el fenómeno de la evaporación a través de técnicas de Aprendizaje Automático (ML, por sus siglas en inglés) debido a los diferentes inconvenientes que se encuentran al momento de estimar la evaporación de manera directa, utilizando otros parámetros climatológicos que presentan relación con este proceso.

Tabla 2. Estadísticos de las variables.

Variable	Máximo		Mínimo		Media		Desviación estandar	
	C	F	C	F	C	F	C	F
T_1	30.40	31.70	17.10	18.20	23.98	25.35	2.97	3.06
T_2	13.80	13.60	-0.70	-4.50	7.84	6.56	4.10	4.98
T_3	22.00	22.20	8.70	8.00	15.93	15.99	3.34	3.79
P	207.80	273.70	0.00	0.00	38.84	36.87	45.69	47.85
HR_1	100.00	99.60	39.90	48.00	80.20	83.44	15.86	13.02
HR_2	60.00	59.30	8.00	7.70	27.43	24.30	12.41	11.01
HR_3	85.80	86.20	20.90	21.30	53.31	53.19	16.86	15.74
RS	1.02e6	9.47e5	5.02e5	4.55e5	7.43e5	7.12e5	1.28e5	1.30e5
V_1	26.50	30.10	13.10	10.40	20.06	20.15	3.18	4.26
V_2	12.90	14.20	4.40	3.90	8.64	7.91	1.87	2.01
EP	350.10	286.10	100.70	54.20	195.15	154.52	63.58	55.97

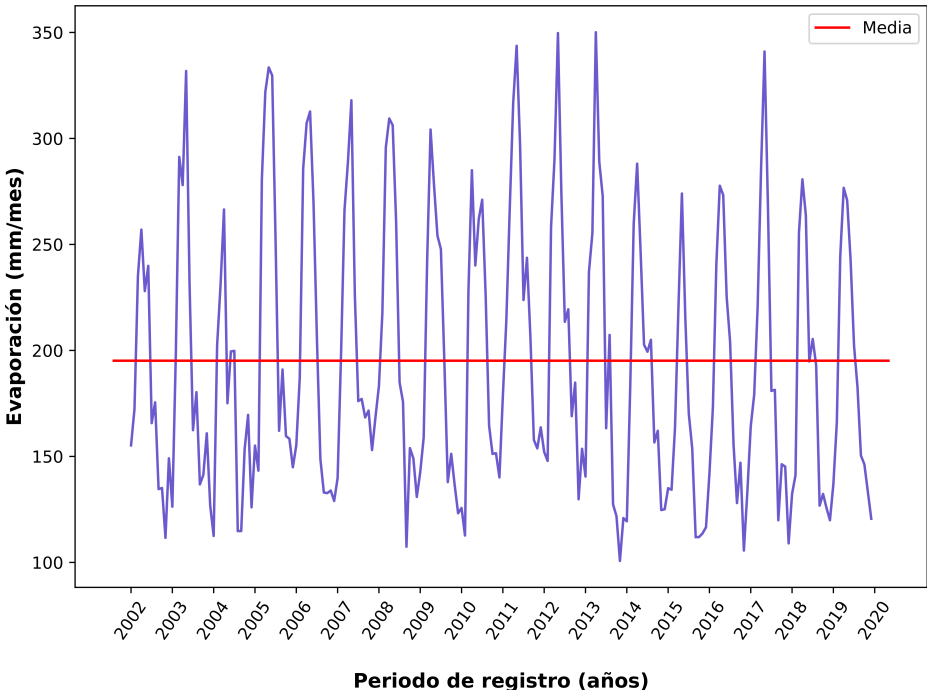

Fig. 2. Comportamiento de la evaporación en la estación: Calera.

Al-Mukhtar [4], realizó una comparación de seis modelos de ML en tres diferentes regiones Bagdad, Basora y Mosul, de Irak, los cuales son regresión condicional de bosques aleatorios (Cforest), regresión spline adaptativa multivariada (MARS), regresión bagged splines multivariada adaptativa (BaggedMARS), modelo de árboles de decisión (M5), k-Vecinos más cercanos (KNN) y k-vecinos más cercanos ponderado (KKNN).

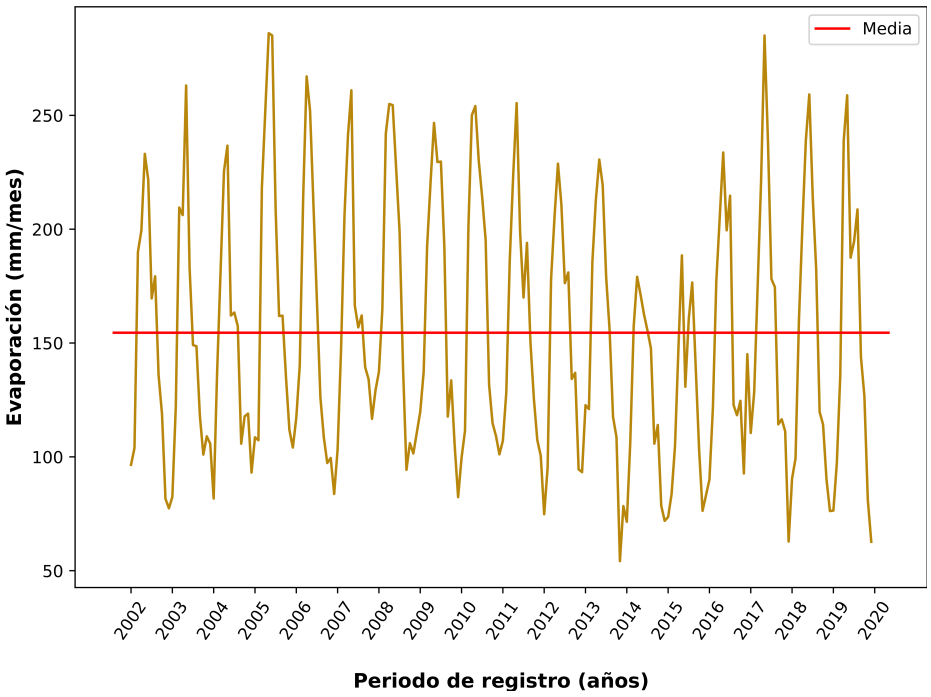

Fig. 3. Comportamiento de la evaporación en la estación: Fresnillo.

Los resultados del análisis dieron a conocer que los modelos KNN y M5 fueron los mejores en términos de capacidad predictiva para modelar las tasas de evaporación, comprobando la buena eficiencia que tienen los modelos de ML.

Shabani [15], implementó 4 diferentes modelos de ML para estimar la evaporación en la Provincia de Golestan, al sureste del mar de Caspian, tales algoritmos son el proceso de regresión Gaussiano (GPR), regresión de máquinas de soporte vectorial (SVR), KNN y bosques aleatorios (RF). Los resultados del estudio indican que bajo las condiciones del análisis el mejor modelo fue el GPR, teniendo ligeramente mejor desempeño que los demás modelos.

El objetivo de esta investigación es realizar una comparación del comportamiento de la evaporación a través de los modelos de ML regresión lineal múltiple (MLR), bosques aleatorios (RF) y k-vecinos más cercanos (KNN), evaluando su desempeño bajo las métricas coeficiente de correlación de Pearson (R), coeficiente de eficiencia Nash-Sutcliffe (NSE), raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE).

2. Zona de estudio

La región Calera se localiza en la porción central del estado de Zacatecas; entre los paralelos $22°41'$ y $23°24'$ de latitud norte y entre los meridianos $102°33'$ y $103°01'$ de longitud oeste, cubriendo una superficie aproximada de 2,226 km^2.

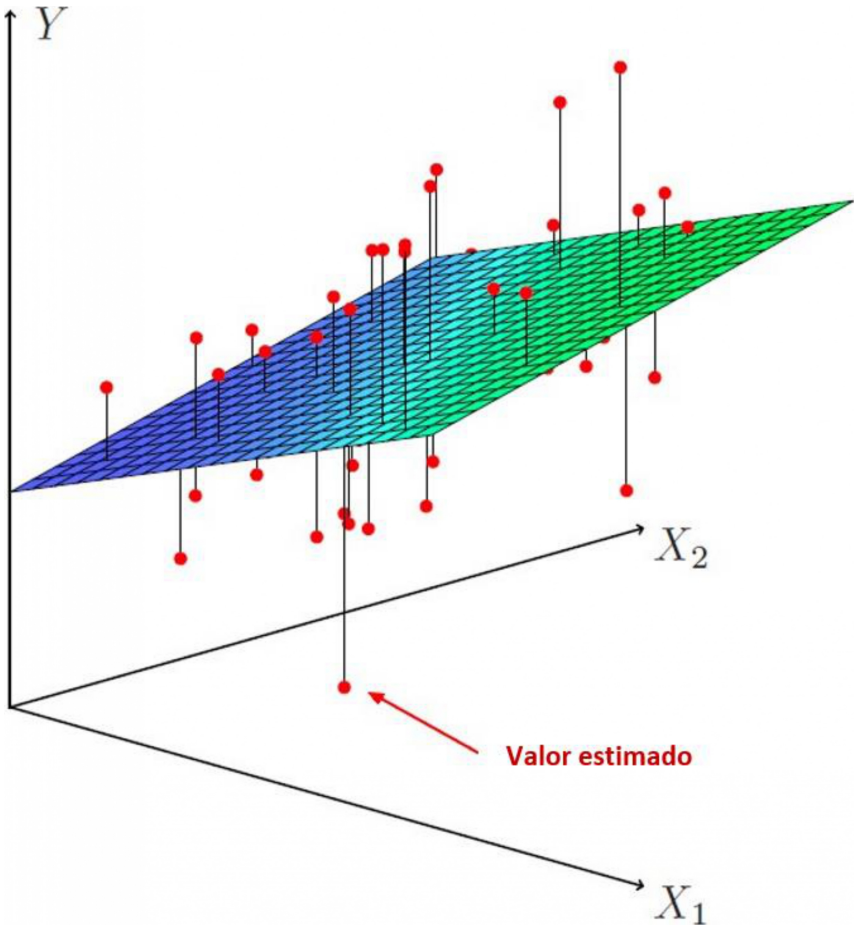

Fig. 4. Regresión lineal múltiple.

De acuerdo con la clasificación de Köppen, modificada por Enriqueta García en 1964 para las condiciones de la República Mexicana [7], en la mayor superficie de la región prevalece el clima semi-seco templado BS1kw, clima seco estepario (BS), que corresponde con el más seco de este tipo de climas, subtipo semi-seco (tipo 1).

Se caracteriza por presentar una temperatura media anual que varía entre $18°C$ y $22°C$, la temperatura media del mes más frío es menor de $18°C$, con invierno fresco y régimen de lluvias en verano. Con los registros obtenidos para el periodo 1980-2009, utilizando el método de isoyetas e isotermas, se determinaron valores de precipitación, temperatura y evaporación potencial media anual de 425 mm, $16.3°C$ y 2,263 mm, respectivamente [5].

Las estaciones climátológicas automáticas analizadas son CEZAC y El Pardillo 3, ubicadas en el municipio de Calera y Fresnillo, así como, las estaciones convencionales ubicadas en estos municipios. En la Tabla 1 se muestran las coordenadas de las estaciones climatológicas.

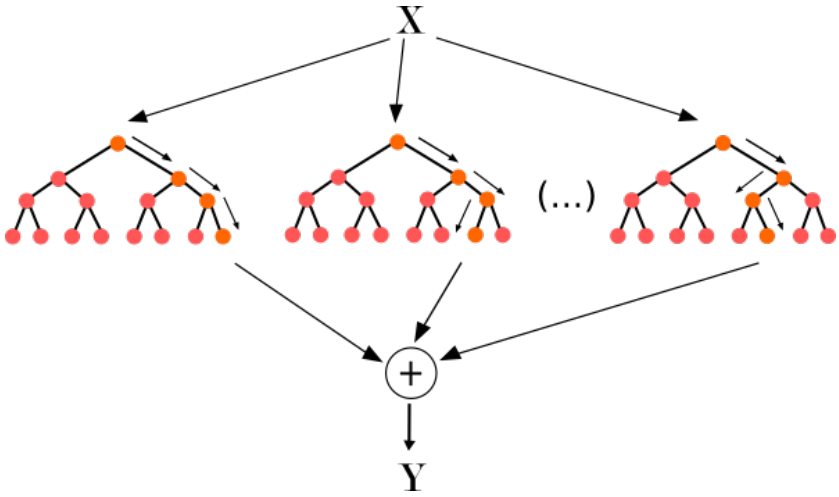

Fig. 5. Bosques aleatorios.

La evaporación se estimó a través del método del tanque evaporímetro Clase A, el cual consiste en un tanque de 120.7 cm de diámetro y 25 cm de profundidad, que debe ser colocado a 5 cm de la superficie del suelo, montado sobre una base de madera que permita la circulación del aire, el material del tanque debe ser de acero resistente a la corrosión, la medición se realiza mediante un micrómetro que indica cuanta cantidad de agua ha evaporado en un intervalo de tiempo y se registra en mm [14]. En la Figura 1 se muestra la ubicación de las estaciones climatológicas.

3. Materiales y métodos

La información climatológica fue proporcionada por el Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP) [9, 14], mediante la Red de Monitoreo Agroclimático del Estado de Zacatecas, las cuales tienen un periodo de registro de 18 años, de enero de 2002 a diciembre de 2019, comprenden los parámetros meteorológicos temperatura media máxima ($T_1, ^\circ C$), temperatura media mínima ($T_2, ^\circ C$), temperatura media ($T_3, ^\circ C$), precipitación (P, mm), humedad relativa media máxima ($HR_1, \%$), humedad relativa media mínima ($HR_2, \%$), humedad relativa media ($HR_3, \%$), radiación solar ($RS, W/\mathrm{m}^2$), velocidad del viento media máxima ($V_1, \mathrm{m}/\mathrm{s}$), y velocidad del viento media ($V_2, \mathrm{m}/\mathrm{s}$).

Por otro lado, mediante la Comisión Nacional del Agua (CONAGUA), a través del Sistema de Información Hidrológica (SIH) se obtuvieron los registros de evaporación (EP, mm) de manera mensual, con el mismo periodo de registro que las estaciones automáticas [6, 14].

En la Tabla 2 se presenta el análisis estadístico de cada variable, para la estación Calera (C) y Fresnillo (F), donde se muestra el valor mínimo, máximo, medio y desviación estandar, en las Figuras 2 y 3 se muestra el comportamiento de la evaporación (EP, mm) a través de los años.

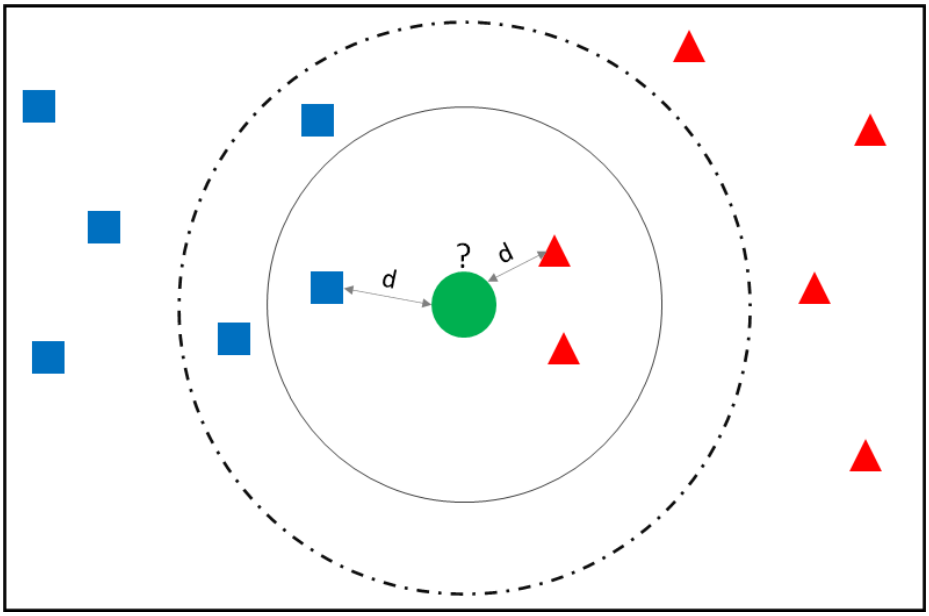

Fig. 6. K-vecinos más cercanos.

3.1. Regresión lineal múltiple (MLR)

La regresión lineal múltiple (MLR, por sus siglas en inglés), es un modelo que ha sido ampliamente utilizado como una herramienta fundamental para modelar relaciones entre variables de manera lineal, y ha sido utilizada en diferentes estudios [3]. MLR establece una relación cuantitativa entre las variables dependientes e independientes, puede ser definido de la siguiente manera:

$$\hat{Y} = b_o + \sum_{i=1}^{n} b_i x_i, \tag{1}$$

donde $\hat{Y}$ es el valor predicho o esperado de la variable dependiente, b_o es el valor de $\hat{Y}$ cuando todos los valores de las variables independientes son iguales a cero, $x_i (i = 1, ..., n)$ son los valores de cada uno de los valores muestra, y $b_i (i = 1, ..., n)$ son los coeficientes (pesos) estimados por la regresión, como se muestra en la Ecuación 1, en la Figura 4 se muestra gráficamente la representación de un modelo de regresión lineal múltiple en tres dimensiones.

3.2. Bosques aleatorios (RF)

Bosques aleatorios (RF, por sus siglas en inglés), es un algoritmo de aprendizaje prominente para problemas de clasificación y regresión. Desde un conjunto de datos único, un número diferente de árboles son construidos de forma aleatoria para el proceso inicial de construcción de árboles total.

Tabla 3. Correlación de las variables con la EP.

Variable	Correlación	
	C	**F**
T_1	0.78	0.84
T_2	0.33	0.49
T_3	0.64	0.76
P	**-0.24**	**-0.06**
HR_1	-0.66	-0.52
HR_2	-0.61	-0.42
HR_3	-0.66	-0.50
RS	**0.85**	**0.89**
V_1	0.66	0.52
V_2	0.55	0.52

Tabla 4. Variables seleccionadas.

Variable	MLR		RF		KNN	
	C	**F**	**C**	**F**	**C**	**F**
T_1	✓				✓	✓
T_2	✓	✓	✓		✓	✓
T_3	✓		✓	✓	✓	✓
P	✓			✓		
HR_1		✓	✓	✓		
HR_2		✓				
HR_3	✓		✓	✓	✓	✓
RS	✓	✓	✓	✓		
V_1	✓		✓	✓		✓
V_2					✓	

La predicción de un bosque aleatorio es entonces, el cálculo del promedio general de las predicciones de todos los árboles. En este algoritmo cada árbol se construye utilizando una mestra de tamaño a_n del conjunto de datos, estos datos solo son utilizados para construir la partición de árboles y posteriormente realizar las predicciones.

Una vez que la observación es seleccionada, el algoritmo forma un control de entrenamiento utilizando cuadrículas o un método aleatorio para la búsqueda. En cada celda, un número de variables de prueba es seleccionada, después son escogidos el número de árboles y los nodos máximos. Matemáticamente el modelo RF se puede representar de la siguiente manera:

$$m_{M,n}(x, \theta_1, ..., \theta_M, D_n) = \frac{1}{M} \sum_{m=1}^{M} m_n(x, \theta_M, D_n), \tag{2}$$

donde $m_n(x, \theta_M, D_n)$ es el valor predicho en el punto x dado por el j-ésimo árbol. $\theta_1, ..., \theta_M$ son las variables aleatorias independientes, distribuidas como una variable aleatoria θ, independiente de la muestra D_n. En la Figura 5 se muestra un esquema de cómo se genera la decisión de cada árbol para realizar una decisión final.

Bosques aleatorios no solo otorga confianza al momento de predecir el modelado, también ayuda en la importancia de la medición, la cual puede ser utilizada para reducir las variables sin perder ninguna información importante.

En el método de bosques aleatorios de Breiman, la variable importante es calculada de la siguiente manera: en primer lugar, se calcula la tasa de error original o el error cuadrático medio de cada árbol formado y también se realiza el mismo experimento con los datos originales con una variable permutada.

Después, se toma la diferencia entre estas dos tasas de error, la nueva medida es la diferencia media de los datos generales dividida por el error estándar de estas diferencias. Esta importante medición de la variable podría usarse para seleccionar un subconjunto de las características más importantes que tienen un gran impacto en la predicción de la variable objetivo [12].

Tabla 5. Resultado de métricas de evaluación.

Modelo	Métricas			
	RMSE(mm)	**MAE(mm)**	**NSE**	R
MLRC	**15.97**	12.56	0.93	0.97
MLRF	**20.53**	14.66	0.87	0.94
RFC	19.46	15.23	0.89	0.95
RFF	21.84	15.78	0.85	0.93
KNNC	18.38	13.36	0.90	0.95
KNNF	24.403	18.269	0.82	0.90

3.3. K-vecinos más cercanos (KNN)

El modelo de los k-vecinos más cercanos (KNN, por sus siglas en inglés), es una técnica no paramétrica que ha sido ampliamente utilizada en el campo de la regresión y clasificación de aprendizaje supervisado.

El fundamento del concepto de regresión KNN es estimar las densidades de probabilidad y funciones de regresión a través de los promedios locales ponderados de la función dependiente. Eso se debe lograr junto con la estimación de la probabilidad condicional basada en los k-vecinos más cercanos de la probabilidad condicional del vector x. La función de densidad utilizada por el modelo KNN se estima de la siguiente manera:

$$f_{NN(x)} = \frac{k/n}{V_{k(x)}} = \frac{k/n}{C_d r_k^d(x)}, \tag{3}$$

donde k es el número de vecinos más cercanos, d son las dimensiones del espacio del vector, C_d es el volumen unitario de la esféra en d dimensiones, $r_{k(x)}$ es la distancia Euclidiana hacia el k-ésimo valor del punto más cercano, y $V_{k(x)}$ es el volumen de la esfera d-dimensional con radio $r_{k(x)}$.

La elección de los k patrones en las observaciones son determinadas con base en la probabilidad del vector condicional utilizando la distancia Euclidiana. En el modelo KNN en todas las variables predictoras se asume tener la misma importancia al momento de estimar la probabilidad condicional [4]:

$$\xi_{t,i} = \sqrt{\sum_{1}^{m} \{S_j x_{j,i} - x_{j,t}\}^2}. \tag{4}$$

Para la estimación de la distancia Euclidiana se tiene la Ecuación 4, donde x es un vector con m predictores $x_{j,i}$ y S_j es el factor ponderado de escala para el j-ésimo predictor. Una vez que se estima la distancia Euclidiana para cada vector característico proyectado, se ordena de manera ascendente.

Así, un conjunto de K-NN casos es seleccionado para que un elemento del conjunto de registro en un tiempo t, se asocie el estado histórico más cercano con el vector actual [4]. En la Figura 6 se observa cómo dependiendo de los vecinos más cercanos, el nuevo valor se asocia a determinado conjunto, esto de acuerdo con la distancia Euclidiana y el número de elementos que tienen características similares.

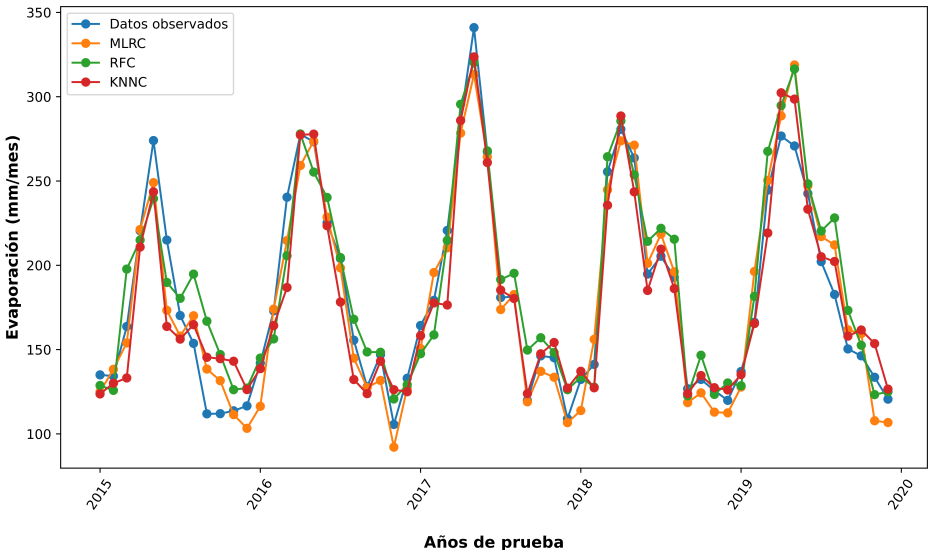

Fig. 7. Modelos estación: Calera.

3.4. Métricas de evaluación

Para evaluar la precisión de los modelos, fueron utilizadas cuatro funciones, coeficiente de correlación de Pearson (R), coeficiente de eficiencia Nash-Sutcliffe (NSE), la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE). Las funciones mencionadas están definidas de la siguiente manera:

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2 \sum_{i=1}^{n}(p_i - \bar{p})^2}}, \tag{5}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - o_i)^2}{n}}, \tag{7}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|p_i - o_i|}{n}, \tag{8}$$

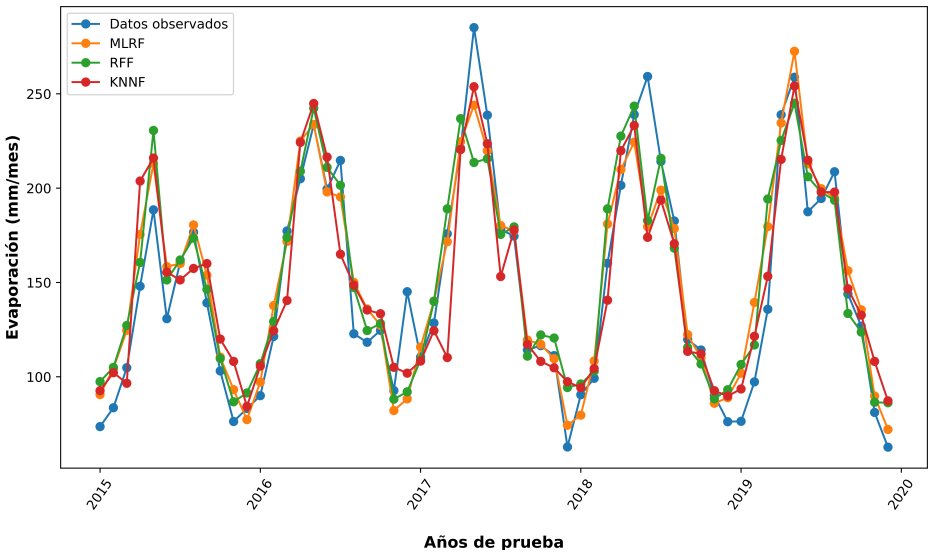

Fig. 8. Modelos estación: Fresnillo.

donde p_i es el i-ésimo valor predicho por los modelos, o_i es el i-ésismo valor observado, y n es un número entero que representa el número total de los datos de la muestra, $\bar{o}$ y $\bar{p}$ son el promedio de los valores observados y predichos, respectivamente. Se utilizó el lenguaje de programación Python en su versión 3.6, en el cual se realizaron los análisis estadísticos, la implementación de los modelos, cálculo de las métricas de evaluación y la elaboración de gráficos.

4. Resultados

Se realizó la comparación de los modelos MLR, RF y KNN, para predecir la EP en dos diferentes estaciones climatológicas, Calera (C) y Fresnillo (F), Zacatecas. A partir de los registros mensuales del periodo 2002-2019, se tomaron los primeros 13 años para entrenamiento y los últimos 5 años para la fase de prueba, con base en aplicar el muestreo aleatorio simple [17].

Se evaluó cada modelo mediante una selección de características exhaustiva, la cual consiste en realizar todas las combinaciones posibles dentro de un conjunto de datos, siempre y cuando la cantidad de variables y muestras lo permitan [11].

Dentro del análisis estadístico, la variable que presentó mayor correlación para ambas estaciones fue la radiación solar ($RS, W/\,\mathrm{m}^2$), con un valor de 0.85 para la estación Calera y 0.89 para Fresnillo, seguido de la temperatura media máxima ($T_1, ^\circ C$) con un valor de 0.78 y 0.84, respectivamente, por otro lado, el parámetro que presentó menor correlación fue la precipitación (PE, mm), con un valor de -0.24 para Calera y -0.06 para Fresnillo, como se muestra en la Tabla 3.

La selección de variables de entrada se realizó tomando como referencia la metrica RMSE, generando los modelos que se muestran en la Tabla 4, los cuales son MLRC, RFC y KNNC, para la estación de Calera, y MLRF, RFF y KNNF, para Fresnillo.

La comparación de los diferentes modelos con respecto a los valores observados, indica una variación espacio-temporal de la evaporación, teniendo un comportamiento diferenciado en primavera-verano y otoño-invierno, los valores con menor ajuste se tienen en la segunda estación (Figuras 7 y 8).

En la Tabla 5 se muestra el resultado de las métricas de evaluación, el mejor rendimiento lo obtuvo el modelo MLR, con R de 0.97 y 0.94, NSE de 0.93 y 0.87, RMSE de 15.97 y 20.53 mm, y MAE de 12.56 y 14.66 mm utilizando como variables climatológicas $T_1, T_2, T_3, P, HR_3, RS$ y V_1; T_2, HR_1, HR_2, y RS, para las estaciones Calera y Fresnillo, respectivamente.

5. Conclusiones

En la presente investigación se realizó una comparación del comportamiento de la evaporación a través de tres modelos de ML, los cuales fueron MLR, RF y KNN, para la región Calera perteneciente al estado de Zacatecas, México. Las técnicas utilizadas bajo las métricas R, NSE, RMSE y MAE, para las estaciones Calera y Fresnillo mostraron que el mejor modelo fue regresión lineal múltiple (MLR), con R de 0.97 y 0.94, NSE de 0.93 y 0.87, RMSE de 15.97 y 20.53 mm, y MAE de 12.56 y 14.66 mm, respectivamente.

La diferencia entre la evaporación estimada y la medida en el tanque se considera significativamente alta, sin embargo, se recomienda para futuros estudios realizar un análisis detallado de la información del tanque e implementación de otras técnicas de ML bajo diversos escenarios, para un mejor comportamiento del proceso de evaporación.

Agradecimientos. Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada mediante la convocatoria “Becas Nacional (Tradicional) 2022 - 1” para la realización de la Maestría en Ciencias del Procesamiento de la Información (MCPI), así como al Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP) y a la Comisión Nacional del Agua (CONAGUA), por la información proporcionada.

Referencias

1. Abed, M., Imteaz, M. A., Ahmed, A. N., Huang, Y. F.: Application of long short-term memory neural network technique for predicting monthly pan evaporation. Scientific Reports, vol. 11, no. 1 (2021) doi: 10.1038/s41598-021-99999-y
2. Aghelpour, P., Bagheri-Khalili, Z., Varshavian, V., Mohammadi, B.: Evaluating three supervised machine learning algorithms (LM, BR, and SCG) for daily pan evaporation estimation in a semi-arid region. Water, vol. 14, no. 21, pp. 3435 (2022) doi: 10.3390/w14213435

3. Al-Ghobari, H. M., El-Marazky, M. S., Dewidar, A. Z., Mattar, M. A.: Prediction of wind drift and evaporation losses from sprinkler irrigation using neural network and multiple regression techniques. Agricultural Water Management, vol. 195, pp. 211–221 (2018) doi: 10.1016/j.agwat.2017.10.005
4. Al-Mukhtar, M.: Modeling of pan evaporation based on the development of machine learning methods. Theoretical and Applied Climatology, vol. 146, no. 3-4, pp. 961–979 (2021) doi: 10.1007/s00704-021-03760-4
5. CONAGUA: Actualización de la disponibilidad media anual de agua en el acuífero Calera (3225) estado de Zacatecas (2020)
6. CONAGUA: Comisión nacional del agua. sih.conagua.gob.mx/ (2020)
7. García, E.: Modificaciones al sistema de clasificación climática de Köppen (2004)
8. Ghazvinian, H., Karami, H., Jun, C., Francis, O., Bateni, S. M., DadrasAjirlou, Y., Band, S.: Laboratory comparison of evaporation rate between Colorado sanken evaporation pan and class a evaporation pan (case study: Semnan, Iran). (2023) doi: 10.20944/preprints202302.0165.v1
9. INIFAP: Instituto nacional de investigaciones forestales, agrícolas y pecuarias. zacatecas.inifap.gob.mx/ (2022)
10. Malik, A., Rai, P., Heddam, S., Kisi, O., Sharafati, A., Salih, S. Q., Al-Ansari, N., Yaseen, Z. M.: Pan evaporation estimation in Uttarakhand and Uttar Pradesh states, India: Validity of an integrative data intelligence model. Atmosphere, vol. 11, no. 6, pp. 553 (2020) doi: 10.3390/atmos11060553
11. Mosre, J., Suárez, F.: Actual evapotranspiration estimates in arid cold regions using machine learning algorithms with in situ and remote sensing data. Water, vol. 13, no. 6, pp. 870 (2021) doi: 10.3390/w13060870
12. Rakhee, Singh, A., Mittal, M., Kumar, A.: Predictive modeling of pan evaporation using random forest algorithm along with features selection. In: 10th International Conference on Cloud Computing, Data Science and Engineering (Confluence), pp. 380–384 (2020) doi: 10.1109/confluence47617.2020.9057856
13. Rezaie-Balf, M., Kisi, O., Chua, L. H. C.: Application of ensemble empirical mode decomposition based on machine learning methodologies in forecasting monthly pan evaporation. Hydrology Research, vol. 50, no. 2, pp. 498–516 (2018) doi: 10.2166/nh.2018.050
14. Secretaría de Economía: Proyecto de norma méxicana. Estaciones meteorológicas, climatológicas e hidrológicas, Parte 3: Condiciones de operación y mantenimiento (2021)
15. Shabani, S., Samadianfard, S., Sattari, M. T., Mosavi, A., Shamshirband, S., Kmet, T., Várkonyi-Kóczy, A. R.: Modeling pan evaporation using Gaussian process regression k-nearest neighbors random forest and support vector machines; comparative analysis. Atmosphere, vol. 11, no. 1, pp. 66 (2020) doi: 10.3390/atmos11010066
16. Sudani, Z. A. A., Salem, G. S. A.: Evaporation rate prediction using advanced machine learning models: A comparative study. Advances in Meteorology, vol. 2022, pp. 1–13 (2022) doi: 10.1155/2022/1433835
17. Verma, S. P.: Estadística básica para el manejo de datos experimentales: Aplicación en la geoquímica: (Geoquimiometría). Universidad Nacional Autónoma de México (2005)
18. Yaseen, Z. M., Al-Juboori, A. M., Beyaztas, U., Al-Ansari, N., Chau, K. W., Qi, C., Ali, M., Salih, S. Q., Shahid, S.: Prediction of evaporation in arid and semi-arid regions: A comparative study using different machine learning models. Engineering Applications of Computational Fluid Mechanics, vol. 14, no. 1, pp. 70–89 (2019) doi: 10.1080/19942060.2019.1680576