ISSN 1870-4069

Evaluación de técnicas de aprendizaje automático supervisado para la predicción de disponibilidad de agua subterránea en acuíferos de México

Alberto González Sánchez[1], Ronald Ernesto Ontiveros Capurata[1], Miguel Antonio Vega Castro[2]

[1] Instituto Mexicano de Tecnología del Agua, Coordinación de Seguridad Hídrica, México

[2] Universidad Politécnica del Estado de Morelos, Maestría en Tecnologías de la Información, México

{alberto_gonzalez,ronald.ontiveros}@tlaloc.imta.mx, 16090507@upemor.edu.mx

Resumen. Hoy en día, la sobreexplotación del agua subterránea es un problema global. En México, hay 653 acuíferos que representan el 39.1% del volumen destinado a usos consuntivos. La Comisión Nacional del Agua es responsable de gestionar este recurso, por lo periódicamente determina la cantidad de agua disponible para extracción considerando factores como el concesionamiento y la recarga. Sin embargo, se ha observado un número creciente de acuíferos en déficit, lo que plantea desafíos para prever su disponibilidad. Los modelos físicos pueden abordar este problema mediante simulaciones, pero requieren una gran cantidad de información, tiempo y recursos. Los algoritmos de aprendizaje supervisado representan una alternativa, ya que pueden detectar tendencias sin el conocimiento profundo que demandan los modelos físicos. Este trabajo evalúa cuatro técnicas para la predicción de disponibilidad de agua en acuíferos: regresión con máquinas de soporte vectorial, árboles de modelos M5', bosques aleatorios (*Random-Forest Regression, RFR*) y redes neuronales artificiales. Los modelos fueron entrenados con información climatológica, uso de suelo y distribución del concesionamiento en años seleccionados entre 1997 y 2015, y fueron evaluados con datos de 2018 y 2020. La comparativa mostró buen desempeño de *RFR*, con un coeficiente *r* alto y errores *RMSE* bajos. La comparativa de acuíferos en déficit mostró una coincidencia del 55.10% para 2018 y del 48.72% para el año 2020. Así, las *RFR* pueden predecir de manera adecuada la disponibilidad de agua en acuíferos a corto plazo, lo que puede ayudar a una gestión más sustentable del recurso.

Palabras clave: Sobreexplotación de acuíferos, *machine learning*, máquinas de soporte vectorial, redes neuronales artificiales, M5', bosques aleatorios.

Assessment of Supervised Machine Learning Techniques for Predicting Groundwater Availability in Aquifers of Mexico

Abstract. Today, the overexploitation of groundwater is a global problem. In Mexico, there are 653 aquifers constituting 39.1% of the volume allocated for consumptive uses. The National Water Commission is responsible for managing this resource, periodically assessing the amount of water available for extraction by considering factors such as concessions and recharge. However, an increasing number of aquifers in deficit have been observed, presenting challenges in accurately predicting their availability. While physical models can address this problem through simulations, they require extensive information, time, and resources. Supervised learning algorithms offer an alternative solution, as they can detect trends without the deep knowledge required by physical models. This work evaluates four techniques for predicting water availability in aquifers: regression with support vector machines, M5' model trees, random forests (Random-Forest Regression, RFR) and artificial neural networks. The models were trained using climatological information, land use and concession distribution data from selected years between 1997 and 2015. They were then evaluated using data from 2018 and 2020. The comparison demonstrated strong performance of RFR, exhibiting a high correlation coefficient (r) and low RMSE errors. The comparison of aquifers in deficit revealed a coincidence of 55.10% for 2018 and 48.72% for 2020. Therefore, RFR can adequately predict the availability of water in aquifers in the short term, aiding in the sustainable management of this vital resource.

Keywords: Aquifers overexploitation, machine learning, support vector machine regression, artificial neural networks, M5', random forests regression.

1. Introducción

Hoy en día, el uso desmedido del agua subterránea ha provocado una reducción en la disponibilidad de este recurso. En México existen 653 acuíferos, que aportan el 39.1% del volumen destinado para usos consuntivos [1], entre los que destacan el consumo humano (14.4%) y la agricultura (60%). La Comisión Nacional del Agua (CONAGUA) ha tratado de hacer un uso eficiente del recurso, estimando a partir del año 2001 la disponibilidad media anual (DMA) de agua por acuífero, considerando el volumen concesionado, la recarga y otras variables [2]. A la fecha, se cuenta con 5 actualizaciones de la DMA de los 653 acuíferos: 2010-2011, 2013, 2015, 2018 y 2020 [3–13]. Históricamente, los valores de DMA han mostrado un deterioro en la cantidad de agua disponible para extracción y un incremento en el número de acuíferos sobreexplotados (Tabla 1).

Un uso sustentable del recurso hídrico y una asignación más equilibrada de los títulos de concesión requieren un análisis complejo que contemple elementos geográficos, ambientales y climatológicos, con el fin de estimar con precisión los valores de DMA.

Tabla 1. Acuíferos en déficit y su disponibilidad promedio según publicaciones oficiales.

Fecha de publicación en el DOF[1]	Cantidad de acuíferos		Disponibilidad promedio total (%)
	En déficit	Con disponibilidad	
08/07/2010, 16/08/2010, 25/01/2011, 14/12/2011[2]	174	479	14.51
20/12/2013	193	460	12.50
20/04/2015	203	450	11.46
04/01/2018	245	408	-2.46
17/09/2020	275	378	-12.01

Los acuíferos tienen una naturaleza dinámica difícil de modelar; responden a cambios en el uso y cobertura del suelo, clima, volumen de recarga y extracción [14]. Predecir la recarga de los acuíferos es complicado, ya que no se puede medir directamente [15]. Un método para detectar el agotamiento de los acuíferos es mediante modelos de simulación física, que demandan gran cantidad de información y son costosos, ya que dependen de la medición directa de variables de campo para su calibración y validación [16]. Una alternativa es utilizar aprendizaje automático (*machine learning, ML*), que es capaz de construir modelos a partir de registros previamente etiquetados [17].

Estos modelos identifican tendencias sin un conocimiento profundo de los atributos subyacentes utilizados en los modelos físicos de flujo de agua subterránea [18]. Diversos algoritmos de *ML* han sido utilizados para abordar el problema de explotación de agua subterránea [19], por ejemplo, redes neuronales artificiales (RNA) [20], bosques aleatorios (*Random Forest*, *RF*) y las máquinas de soporte vectorial (*Support Vector Machines, SVM*) [21], siendo pocos los trabajos realizados en México. En este contexto, este trabajo evalúa el uso de cuatro algoritmos de *ML* (M5', *RF*, RNA y *SVM*) para predecir la disponibilidad de agua en acuíferos de este país.

2. Metodología

2.1 Construcción del conjunto de entrenamiento

El conjunto de entrenamiento se construyó a partir de fuentes de datos oficiales con información histórica de 1997 a 2021 de las variables que afectan la disponibilidad de agua subterránea, como el clima, uso del suelo y distribución de las concesiones para el aprovechamiento del agua (Tabla 2). El volumen disponible por acuífero (en hectómetros cúbicos, hm^3), es la variable de respuesta y fue obtenido de las publicaciones periódicas de la CONAGUA (Tabla 1). La DMA por acuífero en años anteriores al 2011 fue estimada con la información de la fuente FD3 (Tabla 2),

[1] Diario Oficial de la Federación. Órgano del Gobierno Constitucional Mexicano que tiene la función de publicar leyes, reglamentos, acuerdos y demás actos expedidos por los poderes de la Federación.

[2] La primera publicación de la disponibilidad media de los acuíferos se hizo en entregas parciales.

Tabla 2. Fuentes de datos utilizados para la obtención de atributos predictores.

Identificador de Fuente de datos y formato	Descripción
FD1: Uso del suelo y vegetación Formato: vectorial (shapefile)	Capas vectoriales con la clasificación del uso de suelo y vegetación. Series I-VII de uso de suelo de INEGI (1997, 2001, 2005, 2010, 2013, 2016 y 2021). https://www.inegi.org.mx/temas/ususuelo/#descargas
FD2: Clima (temperatura, precipitación y evapotranspiración) Formato: ráster	Capas ráster con los promedios anuales de temperaturas, precipitación y evapotranspiración potencial (1997 a 2021) https://www.globalclimatemonitor.org/.
FD3: Títulos de concesión y sus anexos Formato: tabular (CSV)	Datos de títulos de concesión, tipo de uso (agrícola, industrial, urbano, etc.), volumen amparado y fecha de otorgamiento. https://datos.gob.mx/busca/dataset/concesiones-asignaciones-permisos-otorgados-y-registros-de-obras-situadas-en-zonas-de-libre-alu

utilizando la fecha de registro de la concesión y la diferencia con la recarga reportada en el período 2010-2011. De esta manera se obtiene un rango histórico comparable al de las variables predictoras.

Los datos recopilados se sometieron a un proceso de limpieza e integración. En primer lugar, se verificó la consistencia de las claves, corroborando que fueran homogéneas y que permitieran relacionar las distintas fuentes involucradas.

Cuando fue necesario, se corrigieron datos erróneos y/o atípicos empleando el manejador de base de datos *MySQL*. La integración se realizó considerando la compatibilidad histórica entre los registros, seleccionando como años representativos los correspondientes a las publicaciones de la DMA de los acuíferos, usando 2010 para las entregas realizadas entre 2010-2011 (ver Tabla 1).

Así, se integró un primer conjunto de datos tabular con los atributos año, identificador del acuífero, recarga y DMA (atributo a predecir), al cual se le añadió el resto de los atributos predictores. Para relacionar este conjunto con el uso de suelo, se usó el año representativo para "emparejar" con la serie INEGI más cercana en el tiempo.

De esta forma, 2010 se asoció con la serie IV, publicada entre 2007-2010, 2013 con la serie V, publicada el mismo año, 2015 con la serie VI y 2018 y 2020 con la serie VII. Para contar con más registros históricos, se usó la estimación de disponibilidad realizada con la FD3 en años previos al 2010, asociando estos datos a las series de uso de suelo anteriores.

Así, se hicieron estimaciones para 1997 (Serie 1), 2001 (Serie II) y 2005 (Serie III). En este período no se encontró información para 8 acuíferos (1% del total), por lo que fueron eliminados de todo el conjunto, ajustando el análisis a 645 acuíferos.

Los atributos para clima (temperatura, precipitación y evapotranspiración), se obtuvieron de la FD2. Temperatura y precipitación son las principales variables utilizadas en trabajos similares [19]. La información fue extraída mediante herramientas espaciales de intersección y estadísticas de grupo del software QGIS empleando la capa vectorial de acuíferos.

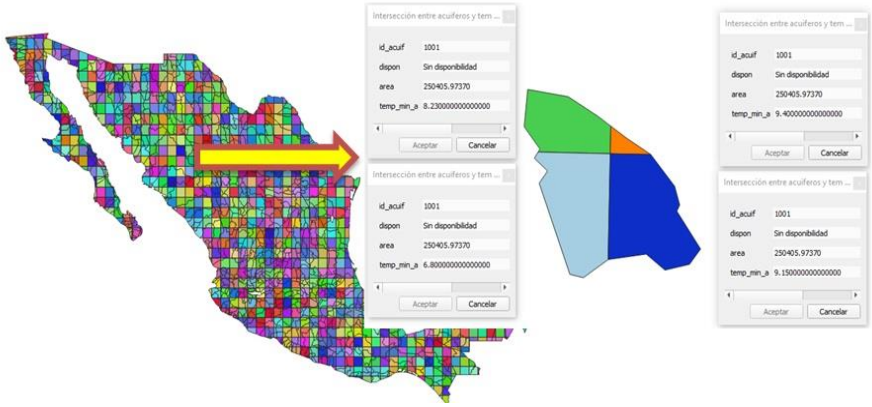

Fig. 1. Procesamiento de extracción de información climática (temperatura) a partir de la intersección del ráster con la capa vectorial de los acuíferos en QGIS.

Tabla 3. Métricas para la evaluación de los modelos. n es el total de observaciones; y_i el valor real de la observación i; $\hat{y}_i$ es el valor estimado por el modelo para i; $\overline{\hat{y}}$ es la media del conjunto de estimaciones; $\overline{y}$: la media del conjunto de observaciones; $r_i = y_i - \hat{y}_i$.

Métrica	Unidades	Cálculo
R	(adim)	$\frac{\sum_{i=1}^{n}(y_i - \bar{y}) \; - \; (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$
RMSE	(misma que el valor estimado y el valor real)	$\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}}$
RRSE	%	$\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} x100$

Por ejemplo, la Figura 1 muestra la extracción del dato de temperatura mínima, donde se muestra el polígono del acuífero 1001 "Valle de Santiaguillo" y los valores de las celdas circunscritas de la capa ráster. Estos valores fueron promediados por acuífero para cada año disponible.

Para obtener la información de un año representativo, se promediaron los datos de los últimos tres años, incluyendo el año de referencia.

Este procedimiento generó un conjunto de entrenamiento con 5160 registros (ocho años por 645 acuíferos) y 23 atributos (incluyendo el atributo a predecir). Los atributos seleccionados fueron los siguientes:

a) Año representativo del período al que corresponden los datos del registro (1997, 2001, 2005, 2010, 2013, 2015, 2018 y 2020) [AÑO].

b) Clave oficial del acuífero (identificador del estado y consecutivo) [ID_ACUIF].

Tabla 4. Valores de métricas obtenidas por M5' en el conjunto de prueba.

Tipo de árbol M5'	Con poda	Métrica		
		R	***RMSE***	***RRSE***
Árbol de modelos	Sí	**0.949**	69.405	39.86%
	No	0.945	74.502	42.78%
Árbol de regresión	Sí	0.000	174.141	100.00%
	No	0.939	**64.848**	**37.24%**

c) Atributos del clima. Temperatura media (°C), precipitación (hm^3) y evapotranspiración potencial (mm) promedio que se presentaron en el acuífero en los últimos tres años [TEMP, PRECIP, ET].

d) Uso de suelo y vegetación. Porcentaje de cada tipo de cobertura de suelo que presenta el acuífero en el año más cercano al representativo [S_AGR, S_ASEN, S_BOSQUE, S_AGUA, S_SELVA, S_VEG, S_OTROS].

e) Tipo de concesionamiento. Porcentaje del uso del agua concesionada para cada tipo de uso estimado para el año representativo [R_ACUA, R_AGRI, R_AGROIND, R_COM, R_DIF_USOS, R_DOMES, R_INDUS, R_OTROS, R_PECUARIO, R_PUB_URBANO, R_SERV].

f) Volumen disponible para extracción que presenta el acuífero (hm^3) (variable a predecir) [VOL_DISP].

2.2 Selección de algoritmos de aprendizaje automático

Se seleccionaron cuatro algoritmos de *ML* comúnmente empleados para predicción numérica; específicamente, las implementaciones de la *suite* para minería de datos Weka [22]. A continuación, se describe cada algoritmo y su parametrización para este trabajo.

Árboles de modelos de regresión lineal M5'. El algoritmo M5', se basa en un árbol de decisión que se construye a partir de un algoritmo recursivo, realizando la toma de decisiones de enrutado en nodos a partir de los valores de los atributos. Al final del enrutado, cada nodo hoja permite obtener el valor de una instancia mediante un modelo de regresión lineal [23], pero también tiene la opción de generar un valor numérico, por lo que en este trabajo se probaron ambas opciones.

También se contempló la opción de podar el árbol, generando cuatro combinaciones posibles: árboles de modelos con poda, árboles de modelos sin poda, árboles de valores constantes con poda y árboles de valores constantes sin poda. Se dejó un mínimo de 2 objetos en cada nodo hoja.

Bosques Aleatorios. La regresión con bosques aleatorios (*Random-Forest Regression, RFR*) tiene su fundamento en el método de *bagging* (embolsado) y los subespacios aleatorios [24]. El algoritmo comienza con la generación de *K* conjuntos obtenidos con la extracción aleatoria de ejemplos con reemplazo del conjunto de aprendizaje, y utiliza cada conjunto para crear un árbol de regresión. En el proceso de

Tabla 5. Valores de métricas obtenidas por *Random-Forest* en el conjunto de prueba.

Tamaño del bosque (árboles)	Límite en profundidad	Variables (*m*)	Métrica		
			R	***RMSE***	***RRSE***
500	10	5	0.959	52.675	30.25%
		8	0.960	54.867	31.51%
	20	5	0.958	53.128	30.51%
		8	0.960	54.752	31.44%
	Sin limitar	5	0.958	53.076	30.48%
		8	0.960	54.752	31.44%
1000	10	5	**0.961**	50.938	29.25%
		8	0.960	54.582	31.34%
	20	5	**0.961**	50.750	29.14%
		8	**0.961**	54.165	31.10%
	Sin limitar	5	**0.961**	**50.681**	**29.10%**
		8	**0.961**	54.126	31.08%

construcción de cada árbol, cada partición es producto de considerar un pequeño conjunto de las variables de entrada de forma aleatoria [25], eligiendo para dividir a la variable con el índice de Gini más bajo. Para la tarea de regresión, el resultado es el promedio de la estimación de los *K* árboles aleatorios del bosque.

De acuerdo con [26], los hiperparámetros más relevantes son el número de variables candidatas por partición (*m*) y el número de árboles (*K*). Los mismos autores sugieren un valor de *m*=*p*/3 para problemas de regresión (*p* es el número de atributos predictores, 23 en este caso), mientras que Weka utiliza *int* (log2(p) + 1)). En este trabajo se consideraron ambas opciones (5 y 8).

Estos autores también sugieren un valor de 500 o 1000 para el número de árboles. En adición, *Weka* permite especificar la profundidad máxima de los árboles, usando en este caso 5, 10 y sin límite. Por lo anterior, para esta técnica se validaron 12 combinaciones de parámetros: *m*={5,8}, *K*={500,1000} y una profundidad de árboles = {5,10,ilimitada}.

Máquinas de Soporte Vectorial para Regresión. La regresión con máquinas de soporte vectorial (*Support Vector Machines Regression, SVMR*) pertenece a un grupo de algoritmos de aprendizaje estadístico supervisado. En su forma más simple, el objetivo de la técnica es obtener una función lineal f(x)=<w,x>+b con $w \in R^N$ y $b \in R$ para un conjunto de entrenamiento $\{(x_1, y_1),...,(x_m, y_m)$. La función *f(x)* debería tener como máximo una desviación ε de los valores y_1 actuales y a la vez ser lo más plana posible. La "planitud" se puede obtener con un valor pequeño para w. El problema de optimización se puede escribir como se muestra en (1) [27]:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)\,, \tag{1}$$

Tabla 6. Valores de métricas obtenidas por las RNAs en el conjunto de prueba.

Ciclos de entrenamiento	Neuronas en la capa oculta	Métrica		
		R	**RMSE**	**RRSE**
1000	5	0.928	91.57	52.58%
	10	**0.930**	**89.63**	**51.47%**
	15	0.924	94.95	54.52%
5000	5	0.921	103.14	59.23%
	10	0.921	102.86	59.07%
	15	0.916	103.47	59.42%
10000	5	0.919	104.84	60.20%
	10	0.919	104.71	60.13%
	15	0.915	104.21	59.84%

Tabla 7. Resultados para las métricas de evaluación (todos los algoritmos).

Algoritmo	Métrica		
	r	***RMSE***	***RRSE***
M5'	0.949	69.405	39.86%
RFR	**0.961**	**50.681**	**29.10%**
RNA	0.930	89.630	51.47%
SVMR	0.920	104.675	60.11%

$$\text{Sujeto a:} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i, \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \tag{2}$$

donde ξ_i y ξ_i^* se introducen como variables de holgura para restricciones inviables.

C se denomina parámetro de regularización y determina la cantidad de desviaciones mayores que ε que son aceptadas. Si no es posible separar el conjunto de ejemplos con una función lineal, se recurre a la transformación del espacio original mediante una función no lineal denominada *kernel*. Para este trabajo, se utiliza la versión de *SVMR* implementada en Weka, que aplica una versión mejorada del algoritmo de aprendizaje de optimización mínima secuencial [28], con C=1 y un *kernel* polinomial de grado 1.

Redes neuronales artificiales (RNA). Las redes neuronales se dividen en una capa de entrada, una de salida y una o más capas ocultas. La capa de entrada consiste de neuronas que reciben las señales o datos del entorno (atributos de entrada). La capa oculta proporciona grados de libertad que le permiten presentar características más complejas. La capa de salida está compuesta de neuronas que proporcionan la respuesta de la red neuronal. Las RNA son empleadas con frecuencia para clasificación y predicción de modelos de series históricas [29].

Existen distintas formas de interconectar las neuronas en una red neuronal (topología). Para este trabajo, se utilizó la topología y esquema de entrenamiento más

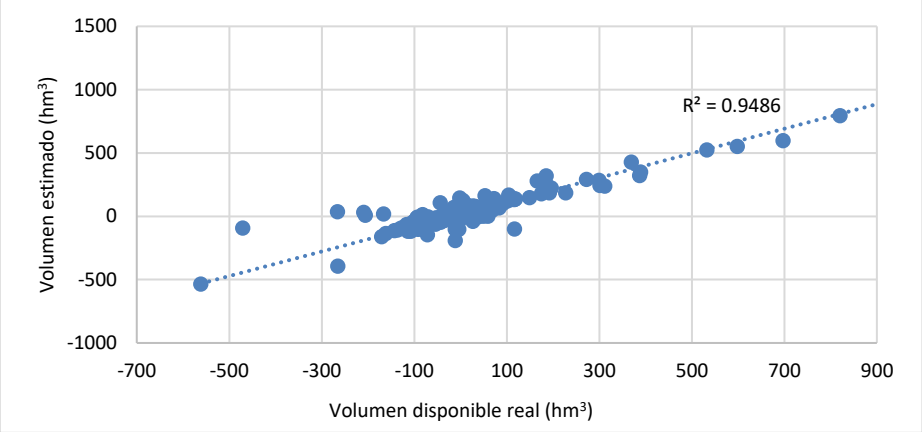

Fig. 2. Volumen disponible real versus el estimado por *RFR* para todos los acuíferos (2018).

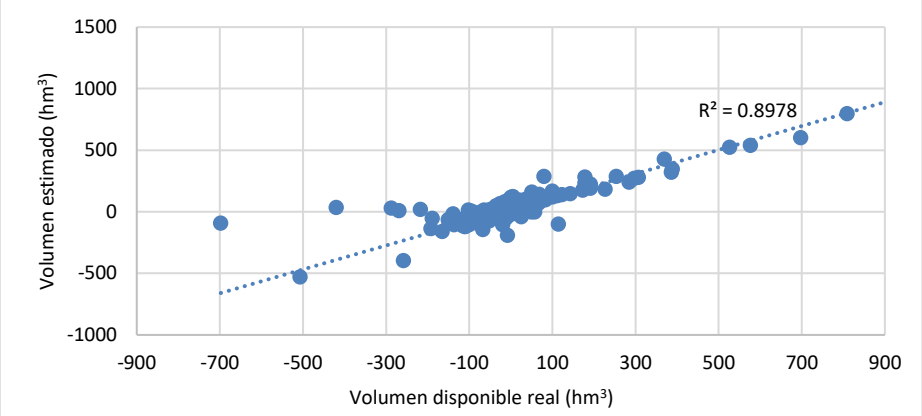

Fig. 3. Volumen disponible real versus el estimado por *RFR* para todos los acuíferos (2020).

común, que es perceptrón multicapa entrenada por retropropagación. Para la capa oculta, se probaron combinaciones de 5, 10 y 15 neuronas, con 1000, 5000 y 10000 ciclos de entrenamiento con decaimiento, ambos parámetros utilizados en trabajos similares realizados con anterioridad [30].

2.3 Evaluación de los algoritmos de aprendizaje

Los algoritmos fueron evaluados usando la técnica *percentage-split*, por lo que el conjunto de aprendizaje se dividió en dos: entrenamiento y prueba. El primer subconjunto se integró con información de los primeros seis años (1997, 2001, 2005, 2009, 2013 y 2015), utilizando 3870 muestras (75% de los registros disponibles). El subconjunto de prueba se integró con la información de los últimos 2 años (2018 y 2020), representando el 25% de los registros restantes.

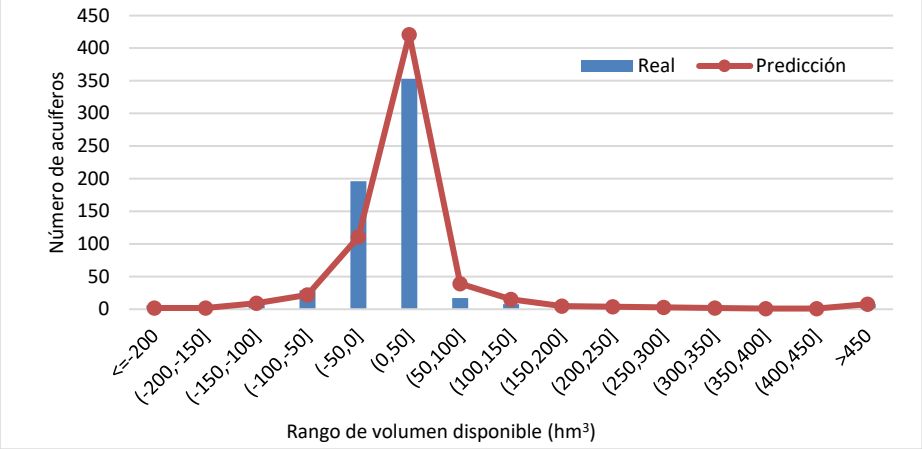

Fig. 4. Histograma de frecuencias de acuíferos por rango de volumen disponible y estimación por *RFR* (2018).

Al tratarse de modelos de predicción numérica, la evaluación del conjunto de prueba se realizó con las métricas de coeficiente de correlación (r), error cuadrático medio (*Root Mean Square Error*, *RMSE*) y el error cuadrático relativo (*Root Relative Square Error*, *RRSE*), que se describen en la Tabla 3 [23].

3. Resultados

En esta sección se muestran los resultados de la evaluación de los algoritmos para el conjunto de prueba (años 2018 y 2020). Primero, se presentan los resultados para cada técnica. Al final, se presenta una comparativa general entre todas las técnicas.

3.1 Resultados por algoritmo

La Tabla 4 muestra los resultados del algoritmo M5'. Se resaltan en negritas los mejores valores de cada métrica. Aunque los árboles de modelos con poda tuvieron una *r* más alta, los errores más bajos se obtuvieron con un árbol de regresión sin poda. La Tabla 5 muestra los resultados del algoritmo *RFR*. Los *RMSE* y *RRSE* más bajos se obtuvieron en un bosque de 1000 árboles, con m=5 y sin límite de profundidad. La Tabla 6 muestra los resultados de las RNAs.

Los mejores valores para las métricas de error se encontraron con 1000 ciclos de entrenamiento y 10 neuronas en la capa oculta. Finalmente, la técnica *SVMR* fue evaluada con los parámetros anteriormente especificados. En este caso, se trató de un único resultado, obteniendo una r=0.920, con un *RMSE*=104.675 y *RRSE*=60.11%.

3.2 Comparación entre algoritmos

La Tabla 7 concentra los mejores resultados encontrados para todos los algoritmos bajo análisis. Se observa que el algoritmo *RFR* obtuvo el valor mayor para *r* y los

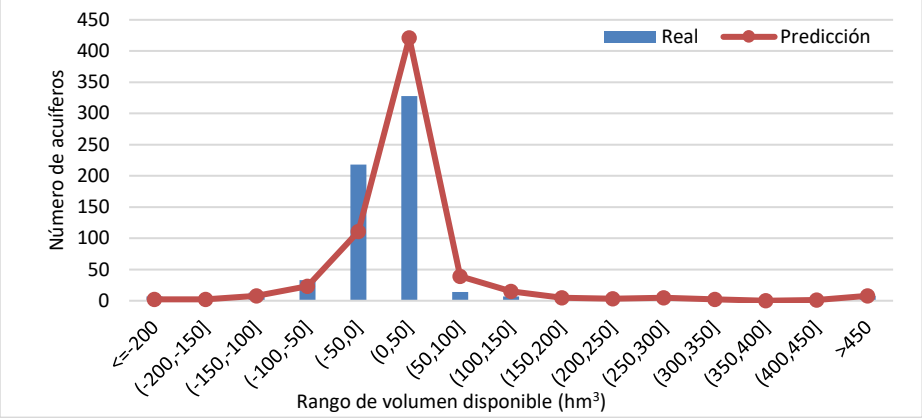

Fig. 5. Histograma de frecuencias de acuíferos por rango de volumen disponible y estimado por *RFR* (2020).

valores más bajos para *RMSE* y *RRSE*. No obstante, se debe considerar que los algoritmos de aprendizaje tienen diversos grados de sensibilidad al peso de sus parámetros. Por ejemplo, *SVM* es más sensible que *RF* [31], y la eficiencia de las RNAs dependen en gran parte de su topología y ciclos de aprendizaje [38].

El uso de alguna técnica de sintonización paramétrica podría mejorar las condiciones de comparación de los algoritmos. Una vez determinado el algoritmo que produce mejores resultados, se puede hacer un análisis más específico.

Así, las gráficas de dispersión de las Figuras 2 y 3 muestran el ajuste obtenido por *RFR* en cada año del conjunto de prueba. Dada la densidad de acuíferos, un análisis de frecuencias puede mejorar la visualización de los errores cometidos por el algoritmo.

Las Figuras 4 y 5 muestran el histograma de los acuíferos por rango de volumen disponible para cada año, sobreponiendo la frecuencia calculada con el volumen estimado por *RFR*. En la comparativa, se observa que la predicción tiene una forma similar a la distribución de probabilidad del valor real de la DMA. Sin embargo, se observa también que el algoritmo subestima la cantidad de acuíferos que tienen una disponibilidad entre el rango de -50 a 0 hm^3, y sobreestima en el rango de 0 a 50 hm^3. Esto es consistente en los dos años presentes en el conjunto de prueba.

Resulta fácil visualizar, que hay una tendencia de los acuíferos que van hacia el estado de disponibilidad bajo 0 (déficit), pero el algoritmo no logra identificar todos los casos. Finalmente, se verificaron las coincidencias entre la predicción de los acuíferos en déficit (disponibilidad negativa) para los años 2018 y 2020, contemplando también su contraparte positiva, obteniendo una coincidencia del 81.24% para 2018 y del 76.79% para 2020. Si la comparación se realiza únicamente con aquellos acuíferos que caen en déficit, la coincidencia es de 55.10% y 48.72%, respectivamente.

4. Conclusiones

En la evaluación de algoritmos de aprendizaje automático para la predicción del volumen disponible en los acuíferos, la regresión con bosques aleatorios (*RFR*) obtuvo

mejores resultados, seguido de M5', RNA y *SVMR*. La ventaja de *RFR* sobre M5' era esperada, ya que el primero se construye a partir de múltiples árboles. Por otra parte, RNA y *SVMR* tienen más combinaciones paramétricas que *RFR*. En este trabajo se utilizaron los valores paramétricos más comunes; sin embargo, los algoritmos de aprendizaje tienen diferentes niveles de sensibilidad al peso de sus parámetros, por lo que una exploración más profunda en el proceso de sintonización (como *grid search*) podría generar una comparación más equitativa.

La clasificación directa de acuíferos que caerán en déficit dada la predicción del volumen disponible realizada por *RFR* tuvo una coincidencia del 55.10% para 2018 y del 48.72% para 2020. La estimación fue consistente, pero baja. En este punto, es importante recordar que el estado de déficit ocurre cuando la disponibilidad es menor a 0, por lo que cantidades cercanas a dicho valor, pero superiores, no se consideran en este estado. Dado que las métricas de *RFR* son buenas, una ligera ampliación del rango para determinar el riesgo de déficit alrededor de la predicción numérica podría aumentar las coincidencias. Desde esta perspectiva, lo más adecuado es usar una técnica de clasificación de dos clases (en déficit/con disponibilidad), por lo que queda pendiente validar este enfoque y su comparación con los resultados de la predicción numérica.

De lo anterior, se concluye que *RFR* puede predecir de manera aceptable la disponibilidad de agua en los acuíferos en corto plazo, no así el estado final de déficit. No obstante, y dado el bajo *RRSE* obtenido, *RFR* puede ser útil para una gestión provisoria del agua subterránea, mejorando la protección y conservación del recurso. En este sentido, es importante señalar que no se sugiere dejar la responsabilidad del concesionamiento a un modelo de aprendizaje. La abstracción inherente al proceso de construcción de estos modelos puede omitir elementos sociales y ambientales importantes, lo que es especialmente crítico en la administración del agua. Depender únicamente de modelos de aprendizaje automático para su gestión plantea un dilema ético; este proceso debe ser transparente, equitativo y responsable, en beneficio de todas las partes involucradas.

Referencias

1. CONAGUA: Estadísticas del agua en México 2018. vol..303 (2018)
2. CONAGUA: Norma oficial mexicana NOM-011-CONAGUA-2000 conservación del recurso agua, que establece las especificaciones y el método para determinar la disponibilidad media anual de las aguas nacionales (2000)
3. CONAGUA: ACUERDO por el que se da a conocer la ubicación geográfica de 371 acuíferos del territorio nacional, se actualiza la disponibilidad media anual de agua subterránea de 282 acuíferos, y se modifica, para su mejor precisión, la descripción geográfica de 202 (2009)
4. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 36 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican (2010)
5. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 44 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican, (2010)

6. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 41 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas que se indican (2010)
7. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 50 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas administrativas que se indican (2011)
8. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 58 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológicas administrativas que se indican (2011)
9. CONAGUA: ACUERDO por el que se da a conocer el resultado de los estudios de disponibilidad media anual de las aguas subterráneas de 142 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican (2011)
10. CONAGUA: ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican (2013)
11. CONAGUA: ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican (2015)
12. CONAGUA: ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las Regiones Hidrológico-Administrativas que se indican (2018)
13. CONAGUA: ACUERDO por el que se actualiza la disponibilidad media anual de agua subterránea de los 653 acuíferos de los Estados Unidos Mexicanos, mismos que forman parte de las regiones hidrológico-administrativas que se indican (2020)
14. Wang, X., Liu, T., Zheng, X., Peng, H., Xin, J., Zhang, B.: Short-term Prediction of Groundwater Level Using Improved Random Forest Regression with a Combination of Random Features. Applied Water Science, vol. 8, no. 5 (2018). DOI: 10.1007/s13201-018-0742-6.
15. Crosbie, R.S., Davies, P., Harrington, N., Lamontagne, S.: Ground Truthing Groundwater-Recharge Estimates Derived from Remotely Sensed Evapotranspiration: A Case in South Australia. Hydrogeology Journal, vol. 23, no. 2, pp. 335–350 (2014). DOI: 10.1007/s10040-014-1200-7.
16. Coulibaly, P., Anctil, F., Aravena, R., Bobée, B.: Artificial Neural Network Modeling of Water Table Depth Fluctuations. Water Resources Research, vol. 37, no. 4, pp. 885–896 (2001). DOI: 10.1029/2000wr900368.
17. Han, J., Kamber, M.: Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann (2006)
18. Steyn, M.: Short-term Stream Flow Forecasting and Downstream Gap Infilling Using Machine Learning Techniques (2018)
19. Uc-Castillo, J.L., Marín-Celestino, A.E., Martínez-Cruz, D.A., Tuxpan-Vargas, J., Ramos-Leal, J.A.: A Systematic Review and Meta-analysis of Groundwater Level Forecasting with Machine Learning Techniques: Current Status and Future Directions. Environmental Modelling & Software, vol. 168, pp. 105788 (2023). DOI: 10.1016/j.envsoft.2023.105788.
20. Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K.: Groundwater Level Forecasting Using Artificial Neural Networks. Journal of Hydrology, vol. 309, no. 1–4, pp. 229–240 (2005). DOI: 10.1016/j.jhydrol.2004.12.001.
21. Kanyama, Y., Ajoodha, R., Seyler, H., Makondo, N., Tutu, H.: Application of Machine Learning Techniques in Forecasting Groundwater Levels in the Grootfontein Aquifer. In:

2nd International Multidisciplinary Information Technology and Engineering Conference, (IMITEC´20), pp. 1–8 (2020). DOI: 10.1109/imitec50163.2020.9334142.
22. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H., Trigg, L.: Weka. In: Proceedings of the Data Mining and Knowledge Discovery Handbook, pp. 1305–1314 (2006). DOI: 10.1007/0-387-25465-X_62.
23. Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W.: Predictive Ability of Machine Learning Methods for Massive Crop Yield Prediction. Spanish Journal of Agricultural Research, vol. 12, no. 2, pp. 313–328 (2014). DOI: 10.5424/sjar/2014122-4439.
24. N., G., Jain, P., Choudhury, A., Dutta, P., Kalita, K., Barsocchi, P.: Random Forest Regression-based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes. Processes, vol. 9, no. 11, pp. 2095 (2021). DOI: 10.3390/pr9112095.
25. Wang, L., Zhou, X., Zhu, X., Dong, Z., Guo, W.: Estimation of Biomass in wheat Using Random Forest Regression Algorithm and Remote Sensing Data. The Crop Journal, vol. 4, no. 3, pp. 212–219 (2016). DOI: 10.1016/j.cj.2016.01.008.
26. Probst, P., Wright, M.N., Boulesteix, A.: Hyperparameters and Tuning Strategies for Random Forest. WIREs Data Mining and Knowledge Discovery, vol. 9, no. 3 (2019). DOI: 10.1002/widm.1301.
27. Vapnik, V., Golowich, S.E., Smola, A.: Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In: Proceedings of the Advances in Neural Information Processing Systems (1997)
28. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K.: Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks, vol. 11, no. 5, pp. 1188–1193 (2000). DOI: 10.1109/72.870050.
29. Maimon, O., Rokach, L.: Introduction to Knowledge Discovery and Data Mining. In: Proceedings of the Data Mining and Knowledge Discovery Handbook, pp. 1–15 (2009). DOI: 10.1007/978-0-387-09823-4_1.
30. Almuhaylan, M.R., Ghumman, A.R., Al-Salamah, I.S., Ahmad, A., Ghazaw, Y.M., Haider, H., Shafiquzzaman, M.: Evaluating the Impacts of Pumping on Aquifer Depletion in Arid Regions Using Modflow, Anfis and Ann. Water, vol. 12, no. 8, pp. 2297 (2020). DOI: 10.3390/w12082297.
31. Fang, P., Zhang, X., Wei, P., Wang, Y., Zhang, H., Liu, F., Zhao, J.: The Classification Performance and Mechanism of Machine Learning Algorithms in Winter wheat Mapping Using Sentinel-2 10 m Resolution Imagery. Applied Sciences, vol. 10, no. 15, pp. 5075 (2020). DOI: 10.3390/app10155075.