DOI: 10.24850/j-tyca-14-02-05
Artículos

Modelado hidrológico basado en el algoritmo KNN: una aplicación para el pronóstico de caudales diarios del río Ramis, Perú

Hydrological modeling based on the KNN algorithm: An application for the forecast of daily flows of the Ramis river, Peru

Efrain Lujano[1], ORCID: https://orcid.org/0000-0002-6543-8324
Rene Lujano[2], ORCID: https://orcid.org/0000-0002-4103-9724
Juan Carlos Huamani[3], ORCID: https://orcid.org/0000-0002-7734-945X
Apolinario Lujano[4], ORCID: https://orcid.org/0000-0002-9386-1613

[1]Escuela Profesional de Ingeniería Agrícola, Universidad Nacional del Altiplano, Puno, Perú / Escuela Profesional de Ingeniería Ambiental, Universidad Peruana Unión, Juliaca, Perú, elujano28@gmail.com
[2]Egresado de la Maestría en Ingeniería de Sistemas, Universidad Nacional del Altiplano, Puno, Perú, rlujano13l@gmail.com
[3]Servicio Nacional de Meteorología e Hidrología (SENAMHI), Lima, Perú, jchc.imf2014@gmail.com
[4]Autoridad Nacional del Agua (ANA), Lima, Perú, apolex23@gmail.com

Autor para correspondencia: Efrain Lujano, elujano28@gmail.com

Resumen

El pronóstico de caudales de un río es de gran importancia para el desarrollo de sistemas de alerta temprana. Los algoritmos de inteligencia artificial han demostrado ser una herramienta eficaz en la modelación hidrológica basado en datos, pues permiten establecer relaciones entre los datos de entrada y salida de una cuenca hidrográfica, y así tomar decisiones basado en datos. Este artículo investiga la aplicabilidad del algoritmo k vecino más cercano (KNN) para el pronóstico de caudales medios diarios del río Ramis en la estación hidrométrica Ramis. Como insumo de entrada al algoritmo de aprendizaje automático KNN utilizamos un conjunto de datos de precipitación media de la cuenca y caudal medio diario de estaciones hidrometeorológicas con varios rezagos. El rendimiento del algoritmo KNN se evaluó cuantitativamente con métricas de habilidad hidrológica, como el error porcentual absoluto medio (MAPE), anomalía del coeficiente de correlación (ACC), eficiencia de Nash-Sutcliffe (NSE), eficiencia de Kling-Gupta (KGE') y ángulo espectral (SA). Los resultados para realizar pronóstico de caudales del río Ramis con el algoritmo de aprendizaje automático KNN alcanzaron altos niveles de confiabilidad, sobre todo con rezagos de caudales de uno y dos días, y precipitación con tres días. El algoritmo utilizado es simple, pero robusto para efectuar pronósticos de caudales a corto plazo, y puede ser integrado como una alternativa para el fortalecimiento del pronóstico hidrológico diario del río Ramis.

Tecnología y ciencias del agua, ISSN 2007-2422, 14(2), 169-203. DOI: 10.24850/j-tyca-14-02-05

Palabras clave: inteligencia artificial, modelado hidrológico basado en datos, aprendizaje automático, río Ramis, k-vecino más cercano.

Abstract

The forecast of river stream flows is of significant importance for the development of early warning systems. Artificial intelligence algorithms have proven to be an effective tool in hydrological modeling data-driven, since they allow establishing relationships between input and output data of a watershed and thus make decisions data-driven. This article investigates the applicability of the k-nearest neighbor (KNN) algorithm for forecasting the mean daily flows of the Ramis river, at the Ramis hydrometric station. As input to the KNN machine learning algorithm, we used a data set of mean basin precipitation and mean daily flow from hydrometeorological stations with various lags. The performance of the KNN algorithm was quantitatively evaluated with hydrological ability metrics such as mean absolute percentage error (MAPE), anomaly correlation coefficient (ACC), Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE') and the spectral angle (SA). The results for forecasting the flows of the Ramis river with the k-nearest neighbor machine learning algorithm reached high levels of reliability with flow lags of one and two days and precipitation with three days. The algorithm used is simple but robust to make short-term flow forecasts and can be integrated as an alternative to strengthen the daily hydrological forecast of the Ramis river.

Keywords: Artificial intelligence, hydrological data-driven modelling, machine learning, Ramis river, K-Nearest Neighbor.

Recibido: 30/11/2020
Aceptado: 29/09/2021

Introducción

Las inundaciones inducidas por el exceso de precipitaciones y desborde de ríos son peligros naturales comunes en las regiones de Perú. La frecuencia con la que se presentan en periodos de avenida provoca pérdidas significativas y daños a la propiedad. Probablemente este fenómeno se vuelva más frecuente con el cambio climático, y los pronósticos confiables y precisos de caudales de un río ayudarían a minimizar los daños asociados con las inundaciones. Los pronósticos precisos a corto plazo (horario y diario) son importantes para predecir inundaciones y desarrollar sistemas de alerta temprana (Mundher, Ahmed, & Abdulmohsin, 2015). Un pronóstico preciso de caudales es crítico para el control óptimo de inundaciones (Solomatine & Xue, 2004).

El uso de modelos basados en procesos se ha convertido en una herramienta esencial para estudiar la respuesta de los regímenes

hidrológicos (Madsen, 2000; Mendez & Calvo-Valverde, 2016), pero una implementación suficientemente representativa y precisa puede llevar a invertir mucho tiempo y costo, además de calibrar una gran cantidad de parámetros. Desde la década de 1930 se han desarrollado numerosos modelos de lluvia escorrentía y todo el proceso físico del ciclo hidrológico se formula matemáticamente en modelos conceptuales que componen gran cantidad de parámetros (Tokar & Johnson, 1999). En un contexto de modelado hidrológico basado en datos "Todos los modelos están equivocados y algunos son útiles"; esta cita es significativa debido a la presencia de diferentes consultas no resueltas y suposiciones deliberadas (Remesan & Mathew, 2015).

Los modelos basados en datos, en especial las técnicas de aprendizaje automático (ML), no requieren ecuaciones físicas complejas y supuestos parámetros que necesitan los modelos basados en procesos. Debido a la simplicidad en su implementación, algoritmos de ML y predicción más precisa se han aplicado de manera amplia en el modelado/pronóstico hidrológico (Mundher *et al*., 2016; Remesan & Mathew, 2015; Solomatine & Xue, 2004), logrando buenos rendimientos inclusive con pequeños conjuntos de datos (Veintimilla-Reyes, Cisneros, & Vanegas, 2016). Los modelos basados en datos o modelos de ML son capaces de realizar pronósticos de lluvia escorrentía incluso para un sistema bastante complejo (Solomatine & Xue, 2004).

El ML es considerado un subcampo de la inteligencia artificial (IA) y se divide en tres clases principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo (Igual & Seguí, 2017). Una

revisión de las técnicas de IA, en específico de algoritmos supervisados de ML han demostrado con éxito su aplicabilidad en la predicción de flujo urbano (Xie *et al*., 2020); predicción de inundaciones (Mosavi, Ozturk, & Chau, 2018; Solomatine & Xue, 2004); pronóstico de caudales diarios (Mundher *et al*., 2015); y modelización y pronóstico de caudales medios mensuales (Laqui, 2010; Lujano, Lujano, Quispe, & Lujano, 2014; Mundher *et al*., 2016). Del mismo modo, también se han aplicado en la previsión de energía eólica basada en datos diarios de velocidad del viento (Demolli, Dokuz, Ecemis, & Gokcek, 2019); estimación de la evapotranspiración (Granata, 2019; Xu *et al*., 2018); predicciones hidro-climatológicas (Thakur, Kalra, Ahmad, & Lamb, 2020); modelado de deslizamientos de tierra (Liu *et al*., 2021); estimación de la evapotranspiración de referencia (Alipour, Yarahmadi, & Mahdavi, 2014; Antonopoulos & Antonopoulos, 2017; Mehdizadeh, 2018); en el modelado de evaluación de riesgo de inundación (Wang *et al*., 2015), así como para modelar la susceptibilidad a deslizamientos inducidos por la lluvia (Dou *et al*., 2019); el modelado lluvia-escorrentía (Tokar & Johnson, 1999); y configuración de relaciones altura-caudal (Jain & Chalisgaonkar, 2000).

En referencia a KNN aplicado a variables de recursos hídricos, encontramos investigaciones aplicadas al pronóstico de precipitaciones (Huang, Lin, Huang, & Xing, 2017), predicciones de conjuntos de múltiples modelos de precipitación y temperatura (Ahmed *et al*., 2020), para la previsión de inundaciones en tiempo real (Liu *et al*., 2020), como modelo generador de clima (Sharif & Burn, 2007), así como para

predicción de la energía eólica (Yesilbudak, Sagiroglu, & Colak, 2017) y la completación de datos (Kowarik & Templ, 2016).

Dado que los algoritmos de ML son un enfoque prometedor, este documento tuvo como objetivo evaluar el algoritmo de aprendizaje automático k vecino más cercano para el pronóstico de caudales del río Ramis, basado en datos conocidos del sistema hidrológico (caudales y precipitación), a fin de contribuir en el desarrollo de sistemas de alerta temprana y fortalecimiento del pronóstico hidrológico.

Materiales y métodos

Área de estudio

La zona en la que se realizó este estudio es la cuenca del río Ramis (14 769.62 km^2), que se extiende desde la estación hidrométrica Ramis hasta la cordillera oriental en el departamento de Puno, Perú (Figura 1) y es la unidad hidrográfica con mayor aporte de caudales al lago navegable más alto del mundo (Titicaca). La altitud de la cuenca está comprendida entre 3 812 y 5 749 metros sobre el nivel del mar (msnm), con una

pendiente promedio de 22 % y una longitud del río principal de unos 321 km. Según la clasificación climática del Perú (SENAMHI, 2020), la cuenca en estudio tiene un tipo de clima predominante lluvioso, con otoño e invierno seco. La precipitación promedio multianual para la cuenca es de 700.1 mm (Fernández, 2017), presentándose mayores acumulados de lluvia en verano (diciembre-febrero), con un otoño e invierno seco que hacen la diferencia al periodo de estiaje. El tipo de cobertura de suelo, según la clasificación anual del programa internacional de geósfera-biosfera (IGBP), disponible en *Google Earth Engine* (GEE), colección de imagen ID MODIS/006/MCD12Q1 (Friedl & Sulla-Menashe, 2015), tiene un 0.01 % de cobertura de árboles; 96.86 % de pastizales dominados por plantas herbáceas (< 2 m); 0.03 % de humedales permanentes; 1.68 % de tierras de cultivo; 0.13 % de tierras urbanas y urbanizadas; 0.01 % de hielo y nieve permanente; 1.25 % de áreas áridas, y 0.02 % de cuerpos de agua.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

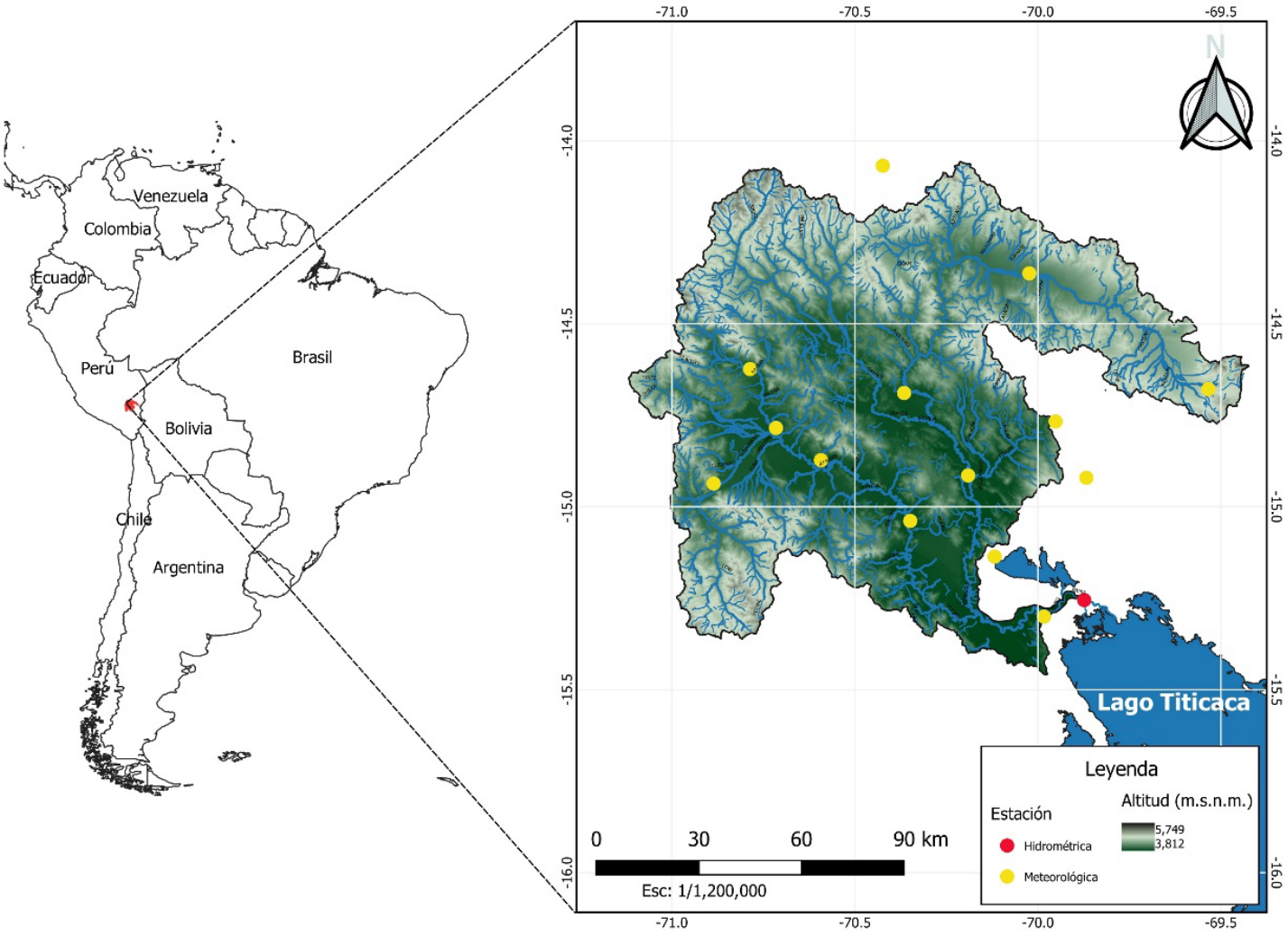

Figura 1. Ubicación del área de estudio.

Las series temporales diarias de precipitación total en milímetros (mm) y caudales medios en metros cúbicos por segundo (m^3/s) se obtuvieron del Servicio Nacional de Meteorología e Hidrología del Perú

(SENAMHI), y el periodo de tiempo utilizado se extiende desde el 01 de septiembre de 2005 al 31 de agosto de 2016. La Figura 1 muestra la ubicación del área de estudio y la distribución espacial de 14 estaciones meteorológicas y una estación hidrométrica.

K vecino más cercanos (KNN)

El algoritmo KNN es uno de los algoritmos más simples en el campo del ML, la idea es memorizar el conjunto de datos de entrenamiento y luego realizar predicciones de cualquier dato nuevo, tomando como referencia los datos de sus vecinos más cercanos en el conjunto de entrenamiento (Shalev-Shwartz, Science, Ben-David, & Science, 2013). Además, KNN es un método no paramétrico que puede usarse como clasificador (Gupta & Mittal, 2018) y regresor (Hossny, Magdi, Soliman, & Hossny, 2020). El algoritmo no asume ningún tipo de ecuación ni relación funcional entre la entrada y la salida (Joshi, 2020):

$$\hat{y} = \frac{\left(\sum_{i=1}^{k} y_i\right)}{k} \qquad (1)$$

donde $\hat{y}$ es el valor de salida; y_i, el i-ésimo vecino más cercano, y k es el número de vecinos más cercanos.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Desarrollo del modelo hidrológico basado en datos

Un paso previo y significativo en algoritmos de ML es la selección de características (variables) más importantes, con el fin de obtener un modelo predictivo más efectivo y evitar características que no contribuyan en el entrenamiento del modelo, y de esta manera disminuir el tiempo de entrenamiento, reducir la complejidad del modelo, y decrecer el sobreajuste. El uso excesivo de una gran cantidad de características en la entrada del modelo conduce a un ajuste perfecto y hace que el modelo memorice el conjunto de datos de entrenamiento y, por lo tanto, pierda la generalización y obtenga resultados pobres en la etapa de validación (Remesan & Mathew, 2015).

Existen varias formas de medir la importancia de las características (Pedregosa, Weiss, & Brucher, 2011; Remesan & Mathew, 2015), pero nos enfocamos en el coeficiente de correlación de Pearson (Tokar & Johnson, 1999) y el algoritmo de importancia de la característica de permutación (Pedregosa *et al*., 2011).

La mejor forma de incorporar características de entrada en un modelo basado en datos es tener en cuenta los retardos de la serie de datos (Mundher *et al*., 2016; Solomatine & Xue, 2004; Tokar & Johnson,

1999), es así que el procedimiento se basa en desarrollar modelos de pronóstico hidrológico que utilizarán memoria, es decir, utilizar valores de caudales y precipitaciones retrospectivos para pronosticar el caudal Q_t del río Ramis.

Entonces, en primera instancia se han seleccionado las entradas en función de un análisis de correlación cruzada entre el conjunto de datos de entrada (precipitación y caudal) con varios rezagos y los caudales de salida Q_t (Remesan & Mathew, 2015; Solomatine & Xue, 2004; Tokar & Johnson, 1999). Aunque la técnica de correlación de Pearson es una técnica adecuada en sistemas lineales, y el proceso de precipitación caudal es no lineal (Remesan & Mathew, 2015), su uso es habitual y popular para seleccionar las entradas apropiadas (Huang & Foo, 2002), pues su fundamento es determinar la fuerza de la relación entre la serie de tiempo de entrada y la serie de tiempo de salida con varios rezagos (Haugh & Box, 1977).

En consecuencia, para inferir qué características tienen mayor impacto en el pronóstico de caudales utilizamos el algoritmo de importancia de permutación, implementado con el modelo predictivo KNN con posibles predictores y la característica predicha. El algoritmo de importancia de permutación es especialmente útil para estimadores no lineales, y se puede calcular en el conjunto de entrenamiento o en el conjunto de prueba o validación extendido cuando los datos son tabulares y una caída de puntuación del modelo es indicativa de cuánto depende el modelo de la característica (Pedregosa *et al*., 2011).

La importancia i_j se calcula con:

$$i_j = s - \frac{1}{K}\sum_{K=1}^{K} s_{k,j} \quad (2)$$

El algoritmo de importancia de permutación requiere como entrada el modelo predictivo ajustado y conjunto de datos (entrenamiento o validación). Se calcula el puntaje de referencia s del modelo ajustado con el conjunto de datos (para verificar el rendimiento del modelo, se utiliza la precisión en un clasificador o R^2 para un regresor). Para cada característica j (columna del conjunto de datos), para cada repetición k en $1, \ldots, K$ barajar de forma aleatoria la columna j del conjunto de datos para generar una nueva versión de conjunto de datos, y calcular el puntaje $s_{k,j}$ del modelo ajustado con la nueva versión del conjunto de datos y quedarnos con K casos.

Modelado hidrológico basado en datos

Los datos de entrenamiento deben ser lo suficientemente grandes como para contener las características de la cuenca; por el contrario, un conjunto de datos insuficiente no permitiría al modelo generalizar los patrones en fenómenos físicos (Tokar & Johnson, 1999).

El proceso de precipitación caudal se modeló utilizando el algoritmo KNN, representando el caudal actual del río Q_t, en función de las características más importantes. La selección del conjunto de datos de entrenamiento (calibración) consideró el 70 % (2 808) del total de datos (4 012), mientras que el restante 30 % (1 204) se tomó en cuenta para la etapa de prueba (validación). Rusli, Yudianto y Liu (2015) indican que la etapa de calibración se realiza para comprender la correlación que existe entre los parámetros del modelo y la respuesta hidrológica de la cuenca y, asimismo, lograr la mejor concordancia entre los caudales observados y simulados. Para obtener el mejor modelo (Ecuación (1)) de pronóstico hidrológico se entrenaron y probaron las configuraciones con las características más importantes (uso de diferentes rezagos/variables para modelar Q_t, determinados mediante la matriz de correlación y el algoritmo de importancia de permutación.

Métricas de bondad de ajuste

La efectividad de los modelos se evaluó mediante cinco métricas de bondad de ajuste diferente (Tabla 1): error porcentual absoluto medio (MAPE), coeficiente de correlación de anomalías (ACC), eficiencia de

Nash-Sutcliffe (NSE), eficiencia de KGE' (Kling, Fuchs, & Paulin, 2012) y ángulo espectral (SA).

Tabla 1. Métricas de bondad de ajuste.

Métricas	**Ecuación**[1]	**Valor óptimo**
Error porcentual absoluto medio (MAPE)	$MAPE = \frac{100}{n}\sum_{i=0}^{n}\left\|\frac{S_i - O_i}{O_i}\right\|$	0
Coeficiente de correlación de anomalías (ACC)	$ACC = \frac{1}{n}\frac{(\sum_{i=1}^{n}(S_i - \bar{S})\,(O_i - \bar{O})}{\sigma_O \sigma_S}$	±1
Eficiencia de Nash-Sutcliffe (NSE)	$NSE = 1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(S_i - \bar{O})^2}$	1
Eficiencia de Kling-Gupta (KGE')	$KGE' = \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$ $\beta = \frac{\mu_S}{\mu_O};\ \gamma = \frac{CV_S}{CV_O} = \frac{\sigma_S/\mu_S}{\sigma_O/\mu_O}$	1
Ángulo espectral (SA)	$SA = arcos\left(\frac{\langle S, O\rangle}{\|S\|_2 \|O\|_2}\right)$	0

[1]Variables: S es el valor simulado; $\bar{S}$ es la media de los valores simulados; O es el valor observado; $\bar{O}$ es la media de los valores observados; σ es la desviación estándar en m^3/s; r es el coeficiente de correlación entre el valor simulado y observado (adimensional); β es la razón de sesgo (adimensional); γ es la razón de variabilidad (adimensional); μ es el valor medio en m^3/s; CV es el coeficiente de variación (adimensional), y los subíndices s y o representan valores observados y simulados respectivamente.

MAPE calcula el error porcentual absoluto medio y su rango es 0 % ≤ MAPE ≤ inf, donde 0 % indica menor error porcentual y por el contrario indican un error porcentual mayor en los datos. Así también, ACC es una medida común en la verificación de campos espaciales y mide la correlación entre el patrón de variación de los valores simulados en comparación con los observados, el rango varía -1 ≤ ACC ≤ 1, donde -1 indica una correlación negativa; 0, aleatoriedad completa, y 1 indica una correlación perfecta del patrón de variación de las anomalías. NSE es una métrica que usa el valor medio como punto de referencia y el rango puede variar de –inf < NSE < 1; mientras el valor se acerca a la unidad es mejor. Por otro lado, KGE' (Kling *et al*., 2012) es la versión modificada de KGE (Gupta, Kling, Yilmaz, & Martinez, 2009) propuesta para evitar correlación cruzada entre las relaciones sesgo y variabilidad; el rango puede variar entre -inf < KGE' < 1; valores cercanos a la unidad no indican sesgo. SA es una medida atractiva para ser utilizada en la coincidencia de espectros; mide el ángulo entre los dos vectores en el hiperespacio e indica qué tan bien coincide la forma de la serie simulada y la observada (no la magnitud); su rango varía entre -π/2 ≤ SA < π/2 (π = pi), donde valores cercanos a cero es mejor.

Para este proceso utilizamos la IDE de *Python Jupyter Notebook* y el paquete *Hydrostats*, que contiene las métricas para caracterizar los errores entre las series de tiempo simuladas y observadas (Roberts, Williams, Jackson, Nelson, & Ames, 2018).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Resultados y discusión

Encontramos que la característica más importante para el pronóstico del caudal Q_t es el caudal con rezago Q_{t-1}, con un coeficiente de correlación igual a 0.99. A medida que ampliamos el rezago en Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5} y Q_{t-6}, el coeficiente de correlación disminuye a 0.96, 0.94, 0.91, 0.90 y 0.88, respectivamente (Figura 2). Si analizamos la relación existente entre precipitación y caudal, la serie de caudal Q_t y la serie de precipitación con rezago P_{t-4} tienen mayor correlación (r = 0.54) respecto a P_{t-1}, P_{t-2}, P_{t-3}, P_{t-5} y P_{t-6}, con coeficientes de correlación de 0.39, 0.45, 0.51, 0.53 y 0.52, respectivamente. Si se desarrollara un modelo de pronóstico Q_t en función de Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5} y Q_{t-6} considerando que tienen mayores valores de correlación obtendríamos un modelo complejo en el cual se incluirían características que no contribuyen al entrenamiento del modelo.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

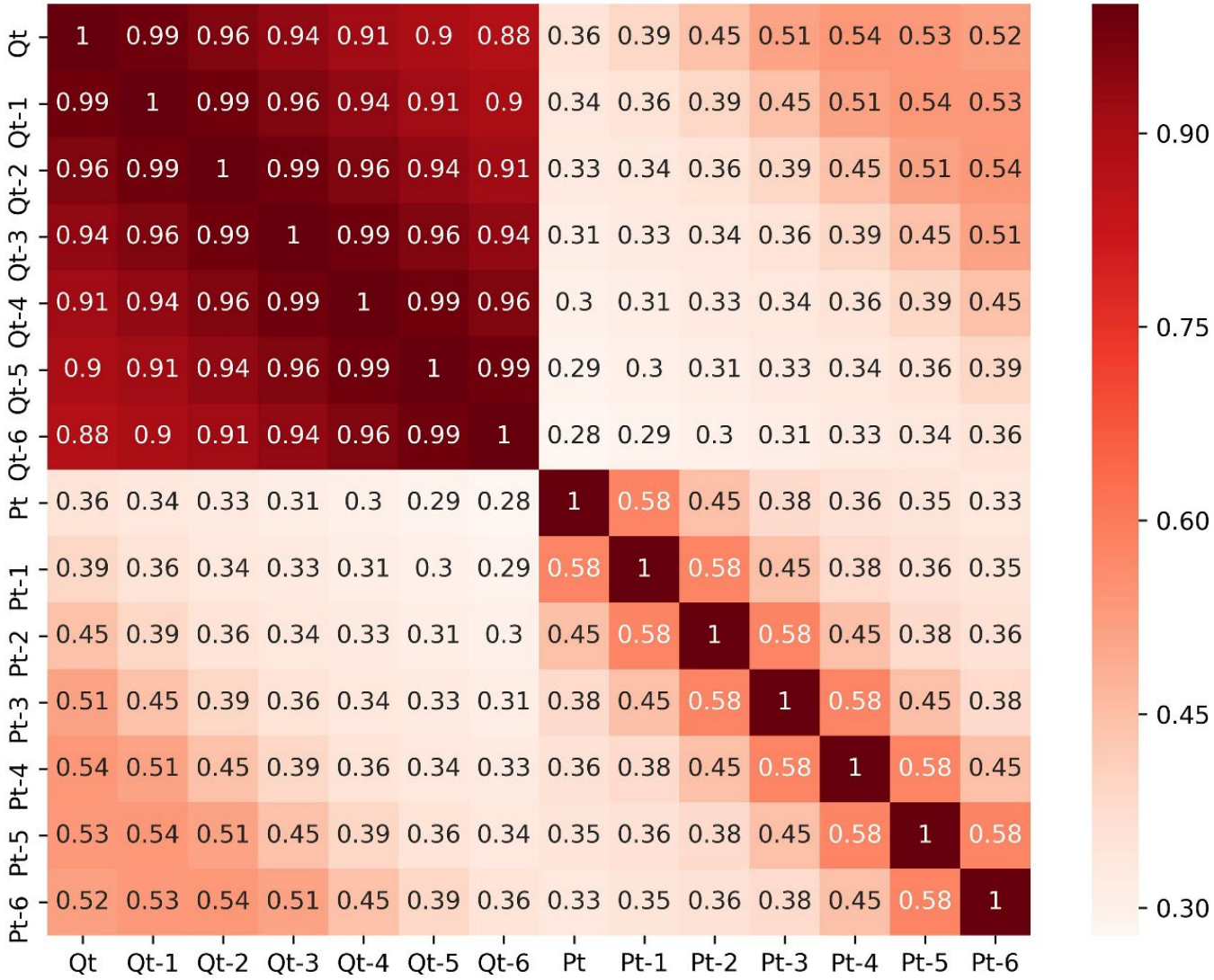

Figura 2. Correlaciones cruzadas de Q_t con rezagos de precipitación y caudal.

Entonces, para una selección definitiva de las características importantes para el modelo, el análisis de correlación se ha complementado mediante la aplicación del algoritmo de importancia de la característica de permutación (Figura 3) y se confirma que el caudal con rezago Q_{t-1} tiene un mayor puntaje (0.779) y es la característica más relevante para pronosticar Q_t, seguido de Q_{t-2} (0.163) y las características con menos importancia Q_{t-3} (0.046), Q_{t-4} (0.039), Q_{t-5} (0.033) y Q_{t-6} (0.037). Así también, la precipitación con rezago P_{t-3} tuvo un mayor puntaje (0.001) respecto a P_t, P_{t-1}, P_{t-2}, P_{t-4}, P_{t-5} y P_{t-6} con menores puntajes de importancia (< 0.001).

Por tanto, el modelo para el pronóstico de caudales se definió mediante una combinación con las características más importantes de precipitación y caudal de acuerdo con los resultados del algoritmo de importancia de la característica de permutación: 1) $Q_t = f(Q_{t-1})$; 2) $Q_t{:}f(Q_{t-1}, Q_{t-2})$; 3) $Q_t{:}f(Q_{t-1}, P_{t-3})$, y 4) $Q_t{:}f(Q_{t-1}, Q_{t-2}, P_{t-3})$, donde Q_t es el caudal a pronosticar, Q_{t-1} y Q_{t-2} son los caudales con rezago de 1 y 2 días, mientras que P_{t-3} es la precipitacion media de la cuenca con rezago de tres días. Aunque P_{t-3} es una característica menos importante respecto a Q_{t-1} y Q_{t-2}, la consideramos como variable de entrada al modelo, pues Pedregosa *et al*. (2011) indica que las características que se consideran de baja importancia, podrían ser muy importantes para un buen modelo y podría incrementar el rendimiento del modelo.

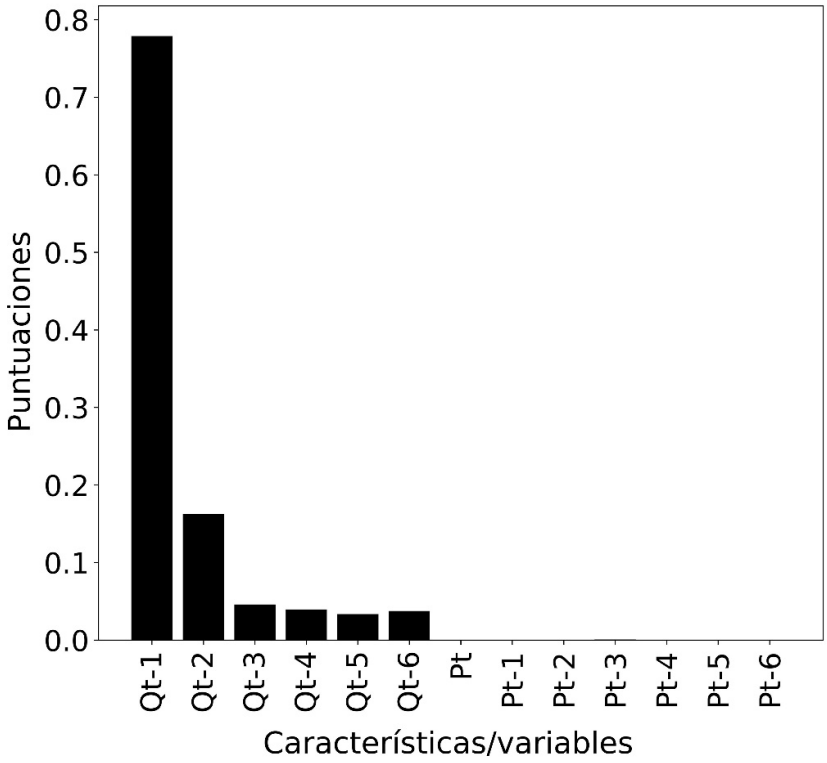

Figura 3. Puntuaciones de importancia de la característica de permutación.

Los resultados de las métricas de bondad de ajuste para el modelo KNN muestran la efectividad para el pronóstico de caudales del río Ramis en la estación hidrométrica Ramis (Tabla 2). Las capacidades predictivas conducen a valores muy altos de NSE (NSE = 0.96) en la etapa de validación, en particular $Q_t{:}f(Q_{t-1}, P_{t-3})$ se caracteriza por valores altos de ACC = 0.989, NSE = 0.979, KGE' = 0.988, valores de error más bajo (MAPE = 6.070 %), y una mejor coincidencia en la forma de la serie simulada y observada (SA = 0.113). $Q_t{:}f(Q_{t-1})$ es el modelo menos

efectivo mostrando valores de ACC = 0.982, NSE = 0.965, KGE' = 0.982, valores de error (MAPE = 7.403 %) y una menor coincidencia en la forma de las series simuladas y observadas (SA = 0.145). También $Q_t{:}f(Q_{t-1}, Q_{t-2}, P_{t-3})$ muestra mejor rendimiento que $Q_t{:}f(Q_{t-1}, Q_{t-2})$ y $Q_t{:}f(Q_{t-1})$, caracterizado por un MAPE = 6.230 %, ACC = 0.987, NSE = 0.975, KGE' = 0.988 y SA = 0.122. Por su parte, $Q_t{:}f(Q_{t-1}, Q_{t-2})$ se identifica por presentar mejores rendimientos respecto a $Q_t{:}f(Q_{t-1})$, con valores de ACC = 0.985, NSE = 0.972, KGE' = 0.985, valores de error (MAPE = 6.546 %), y una similar coincidencia en la forma de la serie simulada y observada (SA = 0.129). Cabe señalar que con un caudal con rezago de un día y precipitación con tres días de rezago, el modelo de pronóstico de caudales muestra un mejor rendimiento respecto a los demás modelos. Esta adición de la P_{t-3} al modelo se corrobora con Pedregosa *et al*. (2011), que aunque sea una característica considerada de baja importancia produce mejores resultados e incrementa el rendimiento del modelo.

Tabla 2. Resultados del rendimiento del algoritmo de KNN – etapa de validación.

Modelo	MAPE (%)	ACC	NSE	KGE'	SA
$Q_t{:}f(Q_{t-1})$	7.403	0.982	0.965	0.982	0.145
$Q_t{:}f(Q_{t-1}, Q_{t-2})$	6.546	0.985	0.972	0.985	0.129
$Q_t{:}f(Q_{t-1}, P_{t-3})$	6.070	0.989	0.979	0.988	0.113
$Q_t{:}f(Q_{t-1}, Q_{t-2}, P_{t-3})$	6.230	0.987	0.975	0.988	0.122

Para continuar con una evaluación adicional de los modelos de pronóstico de caudales, presentamos una serie de diagramas de dispersión (Figura 4), en el que se observa un ajuste casi perfecto entre los valores observados y pronosticados con la línea de 45°, sobre todo para $Q_t: f(Q_{t-1}, P_{t-3})$ (Figura 4c), seguido de los demás modelos (Figura 4a, 4b y 4d). Podemos deducir que el conjunto de datos elegido para entrenar el modelo KNN tiene las mismas propiedades estadísticas y por tanto los parámetros estimados no afecta de modo significativo en el pronóstico de Q_t en el periodo de validación; es así que los valores de MAPE, ACC, NSE, KGE' y SA son similares en el periodo de entrenamiento y validación. Una diferencia significativa en los criterios de evaluación de la bondad de ajuste en el conjunto de entrenamiento y validación podría corresponderse si el modelo se entrena utilizando un conjunto de datos que se desvían enormemente de la situación media y afectan de forma significativa el pronóstico en el periodo de prueba (Antonopoulos & Antonopoulos, 2017). Un conjunto de datos de entrenamiento, validación y prueba con las mismas propiedades estadísticas ayudan a desarrollar el mejor modelo posible (Maier, Jain, Dandy, & Sudheer, 2010).

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

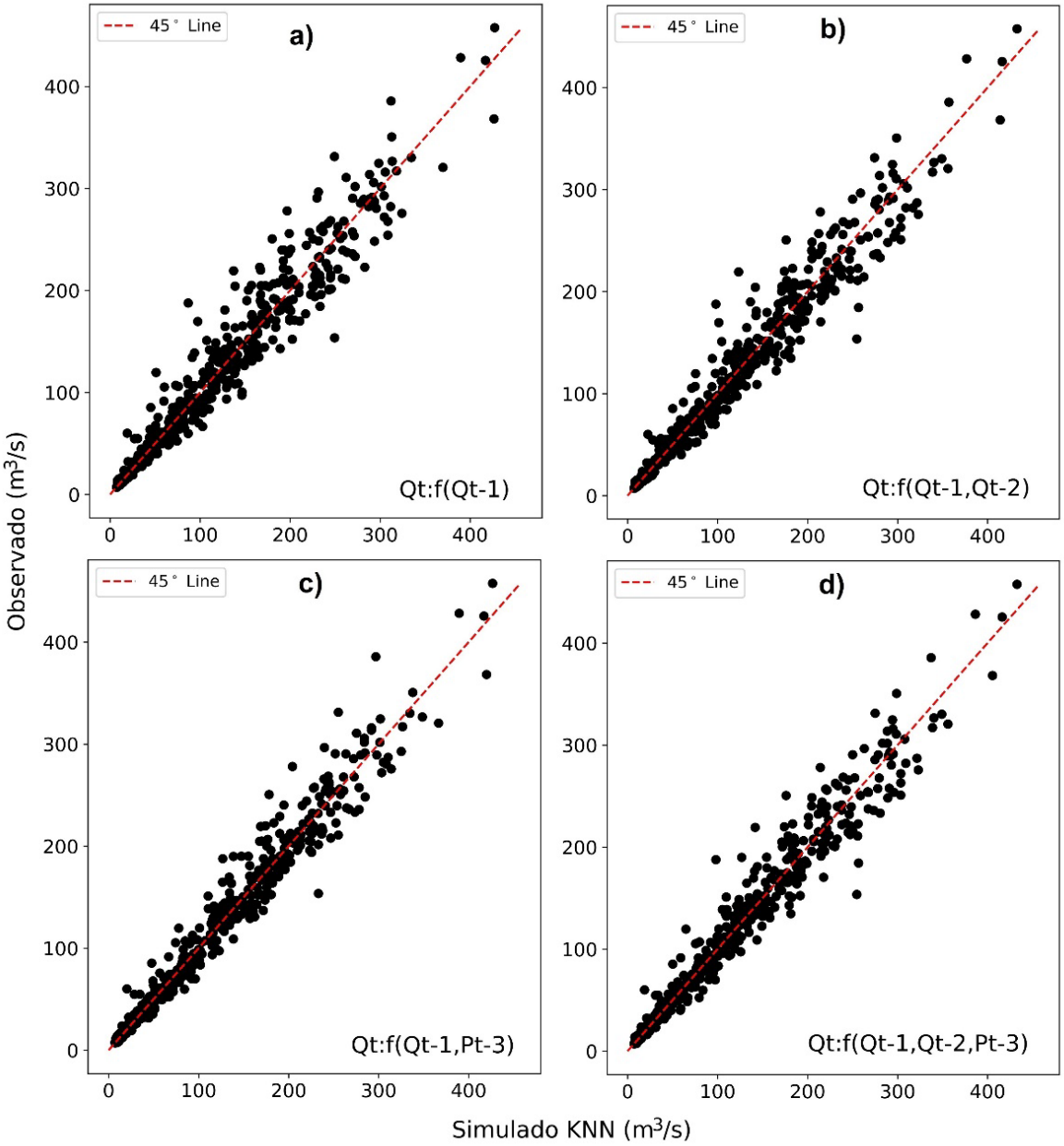

Figura 4. Gráfico de dispersión de caudales simulados y observados.

La Figura 5 ilustra los patrones de caudales observados y simulados con el algoritmo KNN dentro del periodo de validación. Como se observa, los pronósticos basados en datos con KNN pudieron igualar estrechamente a los valores reales. El algoritmo KNN es una herramienta eficaz para el pronóstico de caudales diarios del río Ramis y tiene la ventaja de proporcionar directamente Q_t en función de datos pasados, reduciendo de esta forma inversión en tiempo y costo que se requiere para implementar modelos hidrológicos basados en procesos físicos/conceptuales.

Tecnología y Ciencias del Agua
OPEN ACCESS
Check for updates

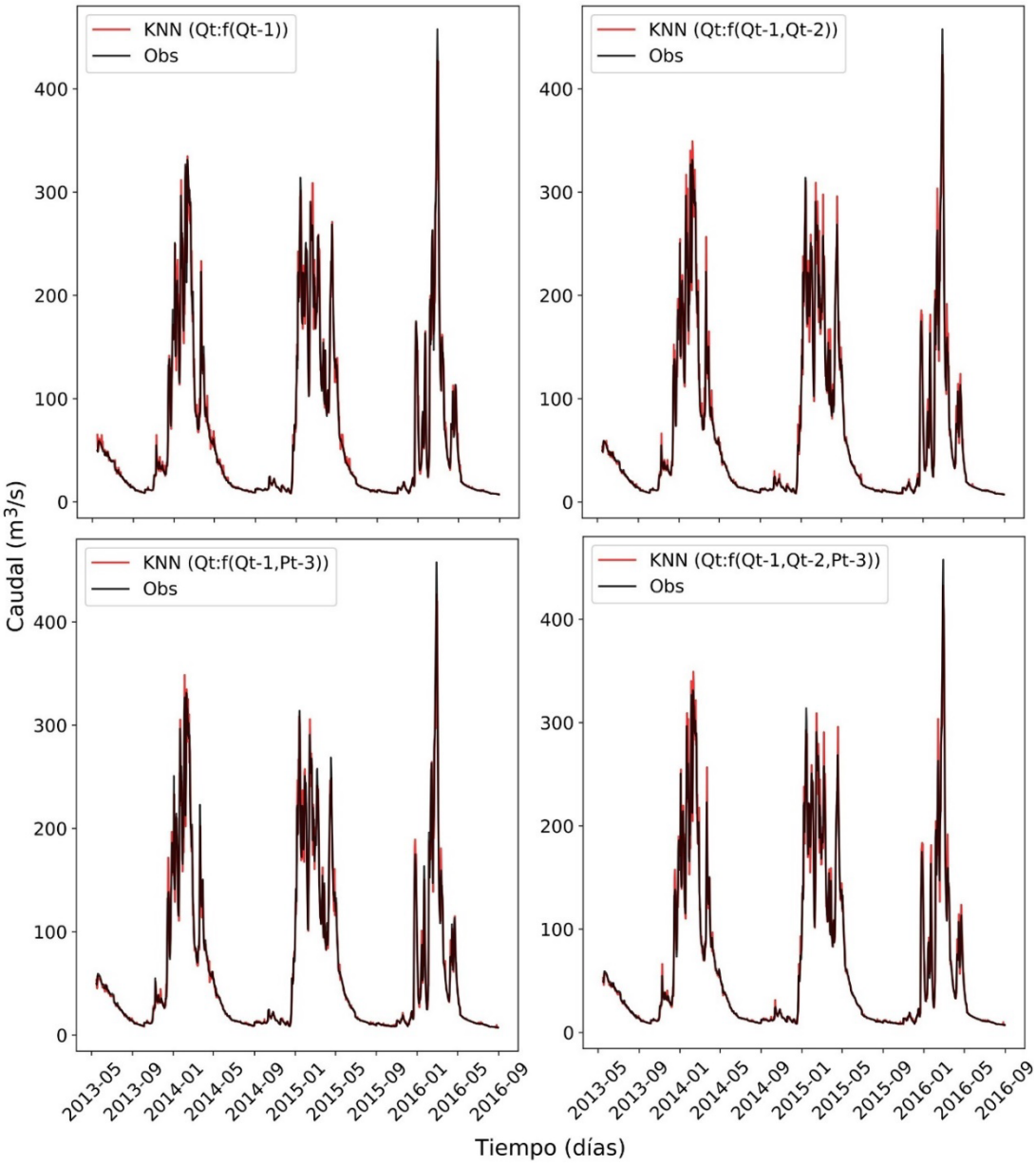

Figura 5. Hidrograma de series de tiempo observados y simulados con el algoritmo KNN-etapa de validación.

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Conclusiones

Este estudio se centró en el modelado hidrológico mediante el uso del algoritmo KNN, explorando su aplicabilidad para el pronóstico de caudales medios diarios del río Ramis. Las características más importantes se seleccionaron en una primera instancia mediante el coeficiente de correlación de Pearson y se complementaron mediante el algoritmo de importancia de las características de permutación. Encontramos que Q_{t-1} es la característica más relevante para el pronóstico de caudales Q_t del río Ramis en la estación hidrométrica Ramis; sin embargo, cuando consideramos Q_{t-1} y P_{t-3} como entrada al modelo, la precisión de KNN se incrementa.

La investigación demuestra que el algoritmo KNN sería un enfoque adecuado para el pronóstico de caudales, pudiendo ser integrado como una alternativa para el fortalecimiento del pronóstico hidrológico diario e implementación en un sistema de alerta temprana.

Agradecimientos

Los autores desean agradecer al Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) por proporcionar la información climática utilizada en esta investigación.

Referencias

Ahmed, K., Sachindra, D. A., Shahid, S., Iqbal, Z., Nawaz, N., & Khan, N. (2020). Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Research*, 236, 104806. Recuperado de https://doi.org/10.1016/j.atmosres.2019.104806

Alipour, A., Yarahmadi, J., & Mahdavi, M. (2014). Comparative study of M5 model tree and artificial neural network in estimating reference evapotranspiration using MODIS products. *Journal of Climatology*, 2014. Recuperado de https://doi.org/10.1155/2014/839205

Antonopoulos, V. Z., & Antonopoulos, A. V. (2017). Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. *Computers and Electronics in Agriculture*, 132, 86-96. Recuperado de https://doi.org/10.1016/j.compag.2016.11.011

Demolli, H., Dokuz, A. S., Ecemis, A., & Gokcek, M. (2019). Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management*, 198(July), 111823. Recuperado de https://doi.org/10.1016/j.enconman.2019.111823

Dou, J., Yunus, A. P., Tien, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y., & Thai, B. (2019). Science of the total environment assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima volcanic island, Japan. *Science of the Total Environment*, 662, 332-346. Recuperado de https://doi.org/10.1016/j.scitotenv.2019.01.221

Fernández, C. (2017). *Modelamiento hidrológico de la región hidrográfica del Titicaca*. Lima, Perú: Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), Dirección de Hidrología.

Friedl, M., & Sulla-Menashe, D. (2015). *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006* [Data set]. NASA EOSDIS Land Processes DAAC. Recuperado de https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD12Q1/

Granata, F. (2019). Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agricultural Water Management*, 217, 303-315. Recuperado de https://doi.org/10.1016/j.agwat.2019.03.015

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80-91.

Gupta, V., & Mittal, M. (2018). KNN and PCA classifier with autoregressive modelling during different ECG signal interpretation. *Procedia Computer Science*, 125, 18-24. Recuperado de https://doi.org/10.1016/j.procs.2017.12.005

Haugh, L. D., & Box, G. E. P. (1977) Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association*, 72(397), 121-130.

Hossny, K., Magdi, S., Soliman, A. Y., & Hossny, A. H. (2020). Detecting explosives by PGNAA using KNN Regressors and decision tree classifier: A proof of concept. *Progress in Nuclear Energy*, 124, 103332. Recuperado de https://doi.org/10.1016/j.pnucene.2020.103332

Huang, M., Lin, R., Huang, S., & Xing, T. (2017). A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Advanced Engineering Informatics*, 33, 89-95. https://doi.org/10.1016/j.aei.2017.05.003

Huang, W., & Foo, S. (2002) Neural network modelling of salinity variation in Apalachicola River. *Water Research*, 36, 356-362.

Igual, L., & Seguí, S. (2017). *Introduction to data science: A python approach to concepts, techniques and applications*. New York, USA: Springer International Publishing.

Jain, S. K., & Chalisgaonkar, D. (2000). Setting up stage-discharge relations using ANN. *Journal of Hydrologic Engineering*, 5, 428-433. Recuperado de https://doi.org/https://doi.org/10.1061/(ASCE)1084-0699(2000)5:4(428)

Joshi, A. (2020). *Machine Learning and Intelligence Artificial* (Springer). Recuperado de https://doi.org/https://doi.org/10.1007/978-3-030-26622-6

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425, 264-277. Recuperado de https://doi.org/10.1016/j.jhydrol.2012.01.011

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7). Recuperado de https://doi.org/10.18637/jss.v074.i07

Laqui, W. F. (2010). Aplicación de redes neuronales artificiales a la modelización y previsión de caudales medios mensuales del río Huancané. *Revista Peruana Geo-Atmosférica RPGA*, 44(2), 30-44.

Liu, M., Huang, Y., Li, Z., Tong, B., Liu, Z., Sun, M., Jiang, F., & Zhang, H. (2020). The applicability of lstm-knn model for real-time flood forecasting in different climate zones in China. *Water (Switzerland)*, 12(2), 1-21. Recuperado de https://doi.org/10.3390/w12020440

Liu, Z., Gilbert, G., Cepeda, J. M., Kydland, A. O., Piciullo, L., Hefre, H., & Lacasse, S. (2021). Modelling of shallow landslides with Machine Learning algorithms. *Geoscience Frontiers*, 12(1), 385-393. Recuperado de https://doi.org/10.1016/j.gsf.2020.04.014

Lujano, E., Lujano, A., Quispe, J. P., & Lujano, R. (2014). Pronóstico de caudales medios mensuales del río Ilave utilizando modelos de redes neuronales artificiales. *Revista de Investigaciones Altoandinas*, 16(1), 89-100.

Madsen, H. (2000). Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of hydrology*, 235(3-4), 276-288.

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling and Software*, 25(8), 891-909. Recuperado de https://doi.org/10.1016/j.envsoft.2010.02.003

Mehdizadeh, S. (2018). Estimation of daily reference evapotranspiration (ETo) using artificial intelligence methods: Offering a new approach for lagged ETo data-based modeling. *Journal of Hydrology*, 559, 794-812. Recuperado de https://doi.org/10.1016/j.jhydrol.2018.02.060

Mendez, M., & Calvo-Valverde, L. (2016). Development of the HBV-TEC Hydrological Model. *Procedia Engineering*, 154, 1116-1123. Recuperado de https://doi.org/10.1016/j.proeng.2016.07.521

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water (Switzerland)*, 10(11), 1-40. Recuperado de https://doi.org/10.3390/w10111536

Mundher, Z., Ahmed, Y., & Abdulmohsin, E. H. (2015). RBFNN *versus* FFNN for daily river flow forecasting at Johor, Malasya. *Neural Computing and Applications*, 27(6). DOI:10.1007/s00521-015-1952-6

Mundher, Z., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., & El-Shafie, A. (2016). Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *Journal of Hydrology*, 542, 603-614. Recuperado de Recuperado de https://doi.org/10.1016/j.jhydrol.2016.09.035

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(2011), 2825-2830.

Remesan, R., & Mathew, J. (2015). Hydrological data driven modelling. In: *Hydrological Data Driven Modelling*. Recuperado de https://doi.org/10.1007/978-3-319-09235-5

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J., & Ames, D. P. (2018). Hydrostats: A Python package for characterizing errors between observed and predicted time series. *Hydrology*, 5(4). Recuperado de https://doi.org/10.3390/hydrology5040066

Rusli, S. R., Yudianto, D., & Liu, J. T. (2015). Effects of temporal variability on HBV model calibration. *Water Science and Engineering*, 8(4), 291-300. Recuperado de https://doi.org/10.1016/j.wse.2015.12.002

SENAMHI, Servicio Nacional de Meteorología e Hidrología. (2020). *Climas del Perú - Mapa de Clasificación Climática Nacional*. Lima, Perú: Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), Dirección de Meteorología y Evaluación Ambiental Atmosférica.

Shalev-Shwartz, S., Science, C., Ben-David, S., & Science, C. (2013). *Understanding machine learning from theory to algorithms*. New York, USA: Cambridge University Press.

Sharif, M., & Burn, D. H. (2007). Improved K-nearest neighbor weather generating model. *Journal of Hydrologic Engineering*, 12(1), 42-51. Recuperado de https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(42)

Solomatine, D. P., & Xue, Y. (2004). M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering*, 9(6), 491-501. Recuperado de https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491)

Thakur, B., Kalra, A., Ahmad, S., & Lamb, K. W. (2020). Bringing statistical learning machines together for hydro-climatological predictions - Case study for Sacramento San Joaquin River Basin, California. *Journal of Hydrology: Regional Studies*, 27, 100651. Recuperado de https://doi.org/10.1016/j.ejrh.2019.100651

Tokar, B. A. S., & Johnson, P. A. (1999). Rainfall-Runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 4(3), 232-239. Recuperado de https://doi.org/https://doi.org/10.1061/(ASCE)1084-0699(1999)4:3(232)

Veintimilla-Reyes, J., Cisneros, F., & Vanegas, P. (2016). Artificial neural networks applied to flow prediction: A use case for the Tomebamba River. *Procedia Engineering*, 162, 153-161. Recuperado de https://doi.org/10.1016/j.proeng.2016.11.031

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527, 1130-1141. Recuperado de https://doi.org/10.1016/j.jhydrol.2015.06.008

Xie, P., Li, T., Liu, J., Du, S., Yang, X., & Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59, 1-12. Recuperado de https://doi.org/10.1016/j.inffus.2020.01.002

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., & Song, L. (2018). Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *Journal of Geophysical Research: Atmospheres*, 123(16), 8674-8690. Recuperado de https://doi.org/10.1029/2018JD028447

Tecnología y Ciencias del Agua

OPEN ACCESS

Check for updates

Yesilbudak, M., Sagiroglu, S., & Colak, I. (2017). A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. *Energy Conversion and Management*, 135, 434-444. Recuperado de https://doi.org/10.1016/j.enconman.2016.12.094