Métodos de inteligencia artificial para la imputación de precipitación pluvial de la cuenca río Ravelo

Campos,J.a, Bellido, J.b Espada,C.c, Huaranca, J.d, Ibarra,A.e

(a) Docente investigador en el área de Hidráulica de la Facultad de Ingeniería Civil (USFX), Destacamento 317 N° 573, Campus Universitario Ex REFISUR, Sucre, Bolivia. E-mail: campos.edgar@usfx.bo.
(b) Docente investigador en el área de Ingeniería de Sistemas de la Facultad de Ingeniería Civil (USFX), Destacamento 317 N° 573, Campus Universitario Ex REFISUR, Sucre, Bolivia. E-mail: bellido.boris@usfx.bo.
(c) Investigadora en el área de Hidráulica de la Facultad de Ingeniería Civil (USFX), Destacamento 317 N° 573, Campus Universitario Ex REFISUR, Sucre, Bolivia. E-mail: carlyveronica93@gmail.com.
(d) Estudiante investigadora en el área de Hidráulica de la Facultad de Ingeniería Civil (USFX), Destacamento 317 N° 573, Campus Universitario Ex REFISUR, Sucre, Bolivia. E-mail: huaranca@gmail.com .
(e) Estudiante investigador en el área de Hidráulica de la Facultad de Ingeniería Civil (USFX), Destacamento 317 N° 573, Campus Universitario Ex REFISUR, Sucre, Bolivia. E-mail: sensey.dumb12@gmail.com

Recibido: 10/10/2024 Aceptado: Publicado:

RESUMEN

En el campo de la hidrología, la Inteligencia Artificial ha tenido una intervención significativa. El objeto es: Evaluar la capacidad de diferentes métodos de inteligencia artificial para la imputación de datos de precipitación en la cuenca del río Ravelo.

Se obtuvieron más de 300.000 datos de precipitación pluvial medidas cada 15 minutos de las gestiones 2019 a las 2023 de la cuenca Río Ravelo en las estaciones de Ravelo, Tumpeka y Cajamarca.

Para determinar las variables usadas se aplicó la correlación de Pearson, obteniendo las variables ambientales para predecir la precipitación pluvial: Presión Barométrica, Humedad relativa, Velocidad del viento y Dirección del viento.

Se definió el conjunto de datos para el entrenamiento, de abril de 2021 a marzo de 2022. Son 140.245 lecturas, Ravelo: 47.331 lecturas, Tumpeka: 46.455 lecturas y Cajamarca: 46.459 lecturas.

Se definió el conjunto de datos de prueba, de mayo 2022 (temporada seca) y diciembre 2022 (temporada de lluvias). Son 17.286 lecturas, Ravelo: 5.762 lecturas, Tumpeka: 5762 lecturas y Cajamarca: 5.762 lecturas.

Los métodos usados son: Regresión lineal, Árbol de decisión (regresor, clasificador) y Regresión Logística.

Se evidenció que todos los métodos usados son capaces de predecir la precipitación pluvial y por ende pueden ser usados en la imputación de la misma. Se observa que los métodos más convenientes para la imputación son la regresión lineal y el árbol de decisión para clasificación, en vista que, según las métricas obtenidas, tiene el mejor desempeño en la predicción.

Palabras clave: Inteligencia Artificial, Regresión lineal, Regresión logística, Árbol de decisión.

ABSTRACT

In the field of hydrology, AI has had a significant intervention. The object is: To evaluate the capacity of different artificial intelligence algorithms for the imputation of precipitation data in the Ravelo River basin. More than 300,000 rainfall data measured every 15 minutes from 2019 to 2023 were obtained from the Ravelo River basin at the Ravelo, Tumpeka and Cajamarca stations. To determine the variables used, the Pearson correlation was applied, obtaining the environmental variables to predict rainfall: Barometric Pressure, Relative Humidity, Wind Speed and Wind Direction. The data set for training was defined, from April 2021 to March 2022. There are 140245 readings, Ravelo: 47331 readings, Tumpeka: 46455 readings and Cajamarca: 46459 readings. The test data set was defined, from May 2022 (dry season) and December 2022 (rainy season). There are 17,286 readings, Ravelo: 5,762 readings, Tumpeka: 5,762 readings and Cajamarca: 5,762 readings. The methods used are: Linear Regression, Decision Trees (regressor, classifier) and Logistic Regression. It was evident that all the methods used are capable of predicting rainfall and therefore can be used in its imputation. It is observed that the most convenient methods for imputation are linear regression and the decision tree for classification, given that, according to the metrics obtained, it has the best performance in prediction.

Keywords: Artificial Intelligence, Linear regression, Logistic regression, Decision tree.

INTRODUCCIÓN

La inteligencia artificial (IA) se refiere al campo de estudio que busca desarrollar sistemas y algoritmos capaces de imitar el comportamiento humano, procesamiento de información y toma de decisiones de manera autónoma (Sansom et al., 2020).

En el campo de la hidrología, la IA ha tenido una intervención significativa, utilizando técnicas como el aprendizaje automático, cómo las que se describen a continuación:

Regresión lineal

Es uno de los modelos más básicos y utilizados en estadística. Se utiliza para analizar la relación entre una variable dependiente (en este caso precipitación) y una o más variables independientes continuas (Montgomery et al., 2019).

La idea principal en la regresión lineal es ajustar una línea recta a los datos de forma tal que minimice la suma de los errores cuadrados entre los valores observados en campo y los valores estimados.

Árbol de decisión

Es un método de aprendizaje supervisado no paramétrico que se utiliza para clasificación y regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo aprendiendo reglas de decisión simples inferidas de las características de los datos. Un árbol puede verse como una aproximación constante por partes.

Los árboles de decisión aprenden de los datos a aproximarse a una curva sinusoidal con un conjunto de reglas de decisión si-entonces-si no. Cuanto más profundo sea el árbol, más complejas serán las reglas de decisión y más ajustado será el modelo (Breiman, 2001).

Un árbol de decisión es una estructura jerárquica que se utiliza en aprendizaje automático de datos para tomar decisiones. Puede ser aplicado a la obtención de datos de precipitación futura puesto que en cada nodo del árbol, se realiza una pregunta sobre una característica específica del conjunto de datos, (en este caso, datos de las diferentes variables ambientales) y se toma una decisión basada en la respuesta a esa pregunta. Esta división se repite en nodos subsiguientes hasta llegar a un nodo hoja, donde se toma una decisión final.

Regresor

Un regresor de árbol de decisión es un tipo de árbol de decisión diseñado específicamente para problemas de regresión. En lugar de devolver una etiqueta de clase en los nodos hoja, los regresores de árbol de decisión devuelven un valor numérico. Para hacer una predicción, el árbol sigue el camino desde la raíz hasta un nodo hoja y devuelve el valor asociado con ese nodo hoja como la predicción. Este método puede ser aplicado en función al conjunto de variables expuestas en la investigación.

Clasificador

El clasificador (DecisionTreeClassifier) se basa en el uso de un árbol de decisiones para tomar decisiones de clasificación. En cada nodo del árbol, se realiza una pregunta sobre una característica específica del conjunto de datos, y según la respuesta, se sigue una rama u otra hasta llegar a una hoja que clasifica el dato. El árbol se construye de manera recursiva dividiendo el conjunto de datos en subconjuntos más pequeños en función de las respuestas a las preguntas (Benzerrouk et al., 2013).

Una de las ventajas del método DecisionTreeClassifier es su capacidad para manejar datos con alta dimensionalidad y características no lineales. Además, permite la interpretación del modelo, ya que se puede visualizar el árbol de decisiones resultante, lo que facilita la comprensión de los factores que influyen en la precipitación.

El método DecisionTreeClassifier es una técnica de aprendizaje automático supervisado que se utiliza para la clasificación de datos. En el campo de la hidrología, este método encuentra aplicaciones en la predicción de la precipitación y en la clasificación de eventos de precipitación en diferentes categorías.

Regresión Logística

Es un método estadístico que permite predecir eventos binarios o categóricos, como la presencia o ausencia de precipitación. Esta técnica se basa en establecer una relación entre la probabilidad de ocurrencia del evento y variables independientes, como la presión atmosférica o la velocidad del viento.

La regresión logística es un método estadístico utilizado para modelar y analizar relaciones entre una variable dependiente binaria (también conocida como variable de respuesta o variable objetivo) y una o más variables independientes.

El uso de la regresión lineal estándar para un resultado de dos niveles puede producir resultados muy insatisfactorios. Es probable que los valores previstos para algunas co-variables estén por encima del nivel superior (normalmente 1) o por debajo del nivel inferior del resultado (normalmente 0). Además, la validez de la regresión lineal depende de que la variabilidad del resultado sea la misma para todos los valores de los predictores. Este supuesto de variabilidad constante no coincide con el comportamiento de un resultado de dos niveles. Por lo

tanto, la regresión lineal no es adecuada para dichos datos y se ha desarrollado la regresión logística para llenar este vacío.

Estaciones meteorológicas

Las estaciones meteorológicas miden diferentes variables ambientales las cuales pueden llegar a influir en la precipitación pluvial.

Variables ambientales

- **Presión Barométrica:**
  La presión barométrica es la medida de la presión atmosférica en un lugar específico. El cambio en la presión barométrica puede influir en la formación de nubes y, por lo tanto, en la precipitación. Las áreas de alta presión tienden a inhibir la formación de nubes y reducir la probabilidad de precipitación, mientras que las áreas de baja presión son propicias para la formación de nubes y la precipitación (Ahrens, 2017).

- **Temperatura:**
  La temperatura también juega un papel importante en la formación y cantidad de precipitación. Durante el proceso de condensación del vapor de agua, la temperatura puede determinar si se formarán nubes de lluvia o de nieve. Las temperaturas más altas favorecen la formación de nubes de lluvia, mientras que las temperaturas más bajas conducen a la formación de nubes de nieve (Muller, 2017).

- **Humedad relativa:**
  La humedad relativa es la cantidad de vapor de agua presente en la atmósfera en relación con la cantidad máxima de vapor de agua que la atmósfera puede contener a una determinada temperatura. Cuanto mayor sea la humedad relativa, mayor será la probabilidad de que ocurra la precipitación, ya que habrá más vapor de agua disponible para la condensación y formación de nubes (Wallace & Hobbs, 2006).

- **Velocidad del viento:**
  La velocidad del viento influye en la precipitación al transportar el vapor de agua a través de la atmósfera. El viento fuerte puede llevar el vapor de agua hacia áreas donde se forman nubes y precipitación, mientras que el viento débil puede impedir la formación de nubes y reducir la probabilidad de precipitación (Anderson, B. M., & Richards, F. J., 2015).

- **Dirección del viento:**
  La dirección del viento puede influir en la distribución espacial de la precipitación. Dependiendo de la dirección del viento predominante, las áreas en el camino del viento pueden recibir más o menos precipitación. Por ejemplo, en las regiones costeras, los vientos que soplan desde el océano pueden llevar aire húmedo y generar precipitación en tierra (Wallace & Hobbs, 2006).

Es importante tener en cuenta que la información proporcionada es general y puede variar dependiendo de las condiciones geográficas y climáticas específicas de cada región.

Objetivo de la investigación

Evaluar la capacidad de diferentes métodos de inteligencia artificial para la imputación de datos de precipitación en la cuenca del río Ravelo.

MATERIALES Y MÉTODOS

Obtención de datos:

Para la investigación se solicitó la información necesaria de la empresa ELAPAS (Empresa Local de Agua Potable y Alcantarillado de SUCRE) gracias a solicitudes y los convenios existentes entre la universidad y la empresa, se obtuvo más de 300.000 datos de precipitación pluvial medidas cada 15 minutos de las gestiones 2019 a las 2023 de la cuenca de Ravelo en las estaciones de Ravelo, Tumpeka y Cajamarca.

Descripción de las variables obtenidas:

- TIMESTAMP: Fecha y hora del registro de la lectura.
- RECORD: Número incremental con el número de la lectura.
- PBar: Medida de la presión barométrica.
- PrecipP: Medición de la precipitación pluvial.
- DirV: Dirección de viento.
- RH: Humedad relativa.
- TA: Temperatura ambiente.
- VelV: Velocidad del viento.
- ET: Evapotranspiración.
- R-Rad: Radiación.
- Rso: No se conoce el significado de la variable.
- Velv_TMn: No se conoce el significado de la variable.

Eliminar las variables que a simple vista no influyen en la precipitación:

- RECORD: El número de lectura no afecta a la precipitación.

- Rso: No se conoce el significado de la variable.
- Velv_TMn: No se conoce el significado de la variable.

Determinación de las variables que más afectan a la precipitación:

- Cálculo de la correlación de Pearson de todas las variables respecto de la precipitación:
  - RH: 0.110408
  - VelV: 0.035822
  - DirV: 0.009436
  - PBar: 0.005318
  - ET: -0.010428
  - TA: -0.028107
  - R-Rad: -0.043143
- Elección de las variables:
  - RH: 0.110408
  - VelV: 0.035822
  - DirV: 0.009436
  - PBar: 0.005318

La correlación evidencia la relación entre la humedad relativa, la velocidad del viento, la dirección del viento, la presión barométrica y la precipitación, tal como indica la teoría.

Definir el conjunto de datos para el entrenamiento, de abril de 2021 a marzo de 2022. Son 140245 lecturas, con el siguiente detalle:

- Ravelo: 47331 lecturas.
- Tumpeka: 46455 lecturas.
- Cajamarca: 46459 lecturas.

Definir el conjunto de datos de prueba, de mayo 2022 (temporada seca) y diciembre 2022 (temporada de lluvias). Son 17286 lecturas, con el siguiente detalle:

- Ravelo: 5762 lecturas.
- Tumpeka: 5762 lecturas.
- Cajamarca: 5762 lecturas.

Se aplicaron cuatro métodos, dos de regresión y dos de clasificación:

- Métodos de regresión:
  - Aplicación de regresión lineal.
  - Aplicación del árbol de decisión para regresión.
- Métodos de clasificación:
  - Agregar una columna binaria: Cuando la precipitación es diferente a cero, el valor uno, en caso contrario es cero.
  - Aplicación de regresión logística.

RESULTADOS Y DISCUSIÓN

Métodos de regresión

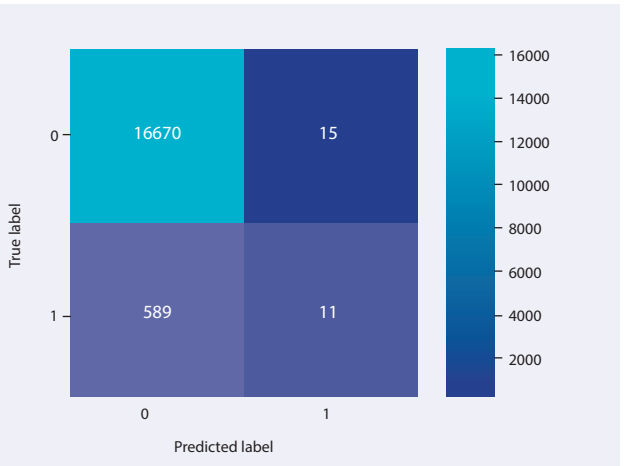
Aplicación de regresión lineal:
Error cuadrático medio (MSE): 0.050305
Coeficiente de Determinación (R2): 0.000276
Aplicación del árbol de decisión para regresión:
Error cuadrático medio (MSE): 0.050463
Coeficiente de Determinación (R2): -0.002852

Métodos de clasificación

Aplicación de regresión logística:
Exactitud (Accuracy), a partir de la matriz de confusión de la Figura 1: 0.965056

Figura 1

Matriz de confusión por el método de Regresión Logística

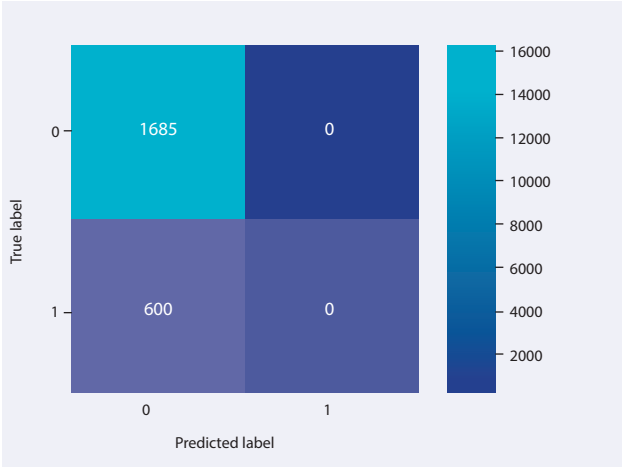


Fuente: Elaboración propia.

Aplicación del árbol de decisión para clasificación:
Exactitud (Accuracy), a partir de la matriz de confusión de la Figura 2: 0.965288

Figura 2

Matriz de confusión por el método de Árbol de Clasificación



Fuente: Elaboración propia.

Con los métodos de regresión; tales como la regresión lineal y el árbol de decisión para regresión, se calculó el error cuadrático medio (MSE) y el coeficiente de determinación (R2), se utilizó ambos parámetros como métricas para evaluar su desempeño.

La similitud de exactitud (accuracy) obtenida por los métodos de clasificación identifica la existencia de métodos recomendables para el relleno de datos de la zona específicamente estudiada.

Analizando los resultados obtenidos, se evidenció que todos los métodos usados son capaces de predecir la precipitación pluvial y por ende puede ser usado en la imputación de la misma. Se observa que los métodos más convenientes para la imputación son la regresión lineal y el árbol de decisión para clasificación, en vista que, según las métricas obtenidas, tiene el mejor desempeño en la predicción.

CONCLUSIONES

Una vez evaluados los métodos, de Inteligencia Artificial, de la regresión lineal, el árbol de decisión y la regresión logística, se puede concluir que todos son capaces de predecir la precipitación pluvial, siendo la regresión lineal y el árbol de decisión los métodos con mejor desempeño en la imputación de datos.

AGRADECIMIENTOS

Al concluir este proyecto de investigación se agradece a la facultad de Ingeniería Civil de la Universidad San Francisco Xavier de Chuquisaca por la enseñanza adquirida, apoyo y motivación por la investigación científica; de igual forma se agradece a la Empresa Local de Agua Potable y Alcantarillado de Sucre (ELAPAS), por la información brindada para la realización y culminación de este proyecto de investigación.

REFERENCIAS BIBLIOGRÁFICAS

- Ahrens, C. D. (2017). Meteorology Today: An Introduction to Weather, Climate, and the Environment. Cengage Learning.
- Anderson, B. M., & Richards, F. J. (2015). Introduction to Physical Meteorology. Academic Press.
- Benzerrouk, H., Nebylov, A., & Salhi, H. (2013). Contribution in Information Signal Processing for Solving State Space Nonlinear Estimation Problems. Journal of Signal and Information Processing, 04(04), 375-384.
- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
- Carleton, A. M., & Henson, R. (2013). Decoding the Weather Machine. W. W. Norton & Company.
- Liou, K. N. (2016). An Introduction to Atmospheric Radiation. Academic Press.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (s. f.).(2019). Introduction to Linear Regression Analysis.
- Muller, R. A., & Thurairajah, B. (2017). Fundamentals of Weather and Climate. Oxford University Press.
- Perez-Paramo, Y. X., Watson, C. J., Chen, G., Thomas, C. E., Adams-Haduch, J., Wang, R., … & Lazarus, P. (2023). Impact of genetic variants in the nicotine metabolism pathway on nicotine metabolite levels in smokers. Cancer Epidemiology, Biomarkers & Prevention, 32(1), 54.
- Prabhakar, G., Raizada, A., & Vishwakarma, B. D. (2019). Evapotranspiration: A Comprehensive Reference (Vol. 6). Springer Nature.
- Sansom, B. J., Bennett, S. J., Atkinson, J. F., & Vaughn, C. C. (2020). Emergent Hydrodynamics and Skimming Flow Over Mussel Covered Beds in Rivers. Water Resources Research, 56(8).
- Wallace, J. M., & Hobbs, P. V. (2006). Atmospheric Science: An Introductory Survey. Elsevier Academic Press.