eib0995885705

# Factores asociados a la citación de artículos biomédicos colombianos: análisis con Machine Learning

Nubia Fernanda Sánchez-Bello*
Jorge Enrique Mejía Quiroga**
Constanza Beatriz Pérez-Martelo**

*Artículo recibido:*
*24 de octubre de 2023*
*Artículo aceptado:*
*28 de febrero de 2024*

*Artículo de investigación*

RESUMEN

Los indicadores de citación pueden medir el impacto o la utilidad de resultados de investigación de un artículo científico, sin embargo, este uso puede ser controversial. Factores intrínsecos y extrínsecos influencian la citación de un artículo, sin mencionar que el comportamiento en las citaciones puede variar entre áreas temáticas, lo cual dificulta las comparaciones entre artículos y disciplinas. Entender que el contexto puede afectar un análisis de citas es esencial para interpretar adecuadamente los indicadores. Por esta razón, buscan reconocerse los factores que inciden en la citación de los artículos de las

* Facultad de Ingeniería y Ciencias Básicas, Universidad Central, Colombia
nsanchezb1@ucentral.edu.co
** Grupo de Investigación "Productividad, Innovación, Desarrollo y Organizaciones", Facultad de Ingeniería y Ciencias Básicas, Universidad Central, Colombia
jmejiaq@ucentral.edu.co cperezm@ucentral.edu.co

revistas biomédicas colombianas indexadas en Scopus a través del uso de algoritmos de Machine Learning. Con los algoritmos 'Gradient Boosting Classifier' y 'Light Gradient Boosting Machine' identificamos características de importancia como el índice *h* del primer y el último autor, acceso abierto, número de autores, palabras clave del artículo, además del número de páginas. Estas características fueron relevantes para el área de interés y pueden brindar un contexto para futuros análisis, considerando que lo relevante de un artículo no debería ser cuántas citaciones atrae, sino si este ayuda a llenar vacíos en el conocimiento.

**Palabras clave:** Análisis de citas; Aprendizaje automático; Investigación biomédica; Colombia

**Factors Associated with Citation of Colombian Biomedical Articles: Analysis with Machine Learning**

*Nubia Fernanda Sánchez-Bello, Jorge Enrique Mejía Quiroga* and *Constanza Beatriz Pérez-Martelo*

## Abstract

Citation indicators can be used to measure the impact or usefulness of research results in a scientific article; however, this usage can be controversial. Intrinsic and extrinsic factors influence the citation of an article, not to mention that citation behavior can differ between thematic areas, which hinders the comparison between articles and disciplines. Understanding that context can affect citation analysis is essential to interpret indicators properly; for this reason, we want to recognize the factors that influence the citation of Colombian biomedical journals indexed in Scopus using Machine Learning algorithms. With 'Gradient Boosting Classifier' and 'Light Gradient Boosting Machine' algorithms, we find characteristics of importance such as the h-index of the first and last author, open access, number of authors and keywords of the article, in addition to identifying the number of pages. These characteristics are relevant to the area of interest and can provide context for future analyses, always considering that what should be relevant about an article is not how many citations it attracts but whether it helps to fill gaps in knowledge.

**Keywords:** Citation Analysis; Machine Learning; Biomedical Research; Colombia

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

## Introducción

La construcción de nuevo conocimiento es uno de los objetivos de las publicaciones científicas; como Cáceres Castellanos (2014) menciona en un editorial: "la ciencia que transmite mejor sus resultados es la más útil" (1). En los artículos científicos, la utilidad suele intentarse medir de forma objetiva con indicadores de citación, partiendo del supuesto de que la citación de un artículo refleja su impacto sobre el conocimiento científico e, incluso, mientras más citas tenga un artículo, más relevante será tanto el artículo como la revista que lo publicó (Martinovich, 2020). Al citar un artículo, se establece un vínculo cuantitativo entre personas, ideas, revistas e instituciones en un contexto temporal que es observable y puede medirse (Mingers y Leydesdorff, 2015).

No toda cita es positiva y existen factores extrínsecos, no relacionados con calidad o contenido, que pueden afectarla (Onodera y Yoshikane, 2015). Dentro de los principales factores de influencia se encuentran su accesibilidad, su diseminación y la autoridad científica de los autores, sin embargo, existen otros que pueden influir como: las publicaciones previas de los autores, la relación del artículo con otros trabajos, las tendencias científicas, la obsolescencia de los resultados, la calidad de los aspectos formales, el contexto teórico del artículo y el tipo de trabajo publicado (Repiso, Moreno-Delgado y Aguaded, 2021).

El uso de indicadores de citación para evaluar la producción científica es frecuente (Ronda-Pupo *et al*., 2022: 111), pero no desprovisto de controversias, ya que asume una relación directa entre relevancia y número de citas desestimando en ocasiones la calidad, mérito, innovación o impacto científico, además, considerarlo referente de calidad en ámbitos de evaluación impulsa a los investigadores a exagerar la importancia de sus hallazgos o a realizar investigaciones poco innovadoras (Stephan, Veugelers y Wang: 2017). Existe también evidencia de que la citación tiene un comportamiento diferencial entre áreas temáticas (Crespo, Li y Ruiz-Castillo, 2012; Crespo, Li y Ruiz-Castillo, 2013; Onodera y Yoshikane, 2015), lo cual dificulta la comparación entre disciplinas.

Se utiliza Machine Learning para identificar relaciones ocultas que afectan la citación, pues existe un problema: debido a la presión por publicar, autores y revistas buscan estrategias para manipular el número de citaciones, por lo cual se requieren modelos que identifiquen anormalidades oportunamente para promover estándares justos de evaluación de la calidad científica (Pradhan, Chakraborty y Nandi, 2019). Su (2020) planteó la citación como una tarea de clasificación binaria basada en características propias de los artículos; con esta estrategia, y utilizando tres algoritmos y una red neuronal, pudo clasificar el 20% de artículos que más citaciones recibirían (103). Otros autores evaluaron el desempeño de modelos de clasificación comparándolos entre sí según su capacidad para predecir el número

de citas de un artículo publicado en Medline, la principal base de datos de artículos biomédicos; con máquinas de soporte vectorial los artículos fueron clasificados según contenido, factor de impacto y conteo de citación demostrando que los modelos diseñados con una tarea específica tienen mejor desempeño que el factor de impacto y el conteo de citación (Aphinyanaphongs, Statnikov y Aliferis: 2006). La información de las citas también se ha analizado con procesamiento de lenguaje natural (Iqbal *et al*., 2021), con estas estrategias se identifica su contexto y contenido para reconocer las razones que motivan una cita. Alohali y su equipo investigaron, en el área de la otología, los factores que influencian el número de citaciones de un artículo científico, utilizando Machine Learning y procesamiento de lenguaje natural encontraron que los resúmenes fueron el elemento que más influía en el número de citas (2022: 10).

Existe una relación entre el número de citas que recibe un artículo y la percepción de su utilidad o relevancia; en este trabajo se quieren presentar los factores que pueden influir sobre la citación para proporcionar un contexto en el análisis de indicadores basados en citas, particularmente al evaluar revistas biomédicas colombianas. Los artículos científicos se han convertido en el principal canal de comunicación para la comunidad de esta área (Navarrete y Pérez, 2019), con todo lo que implica: varios de ellos fueron referentes para la atención de pacientes, en cuestión de días, al ser publicados en revistas consideradas de alta calidad y posteriormente desestimados por fallas en su elaboración (Anderson, Nugent y Peterson, 2021).

Por su rápido avance y relevancia para la salud pública resulta de interés analizar cuáles factores pueden impactar en la citación de un artículo en esta área de conocimiento. El objetivo de este trabajo fue analizar, por medio de algoritmos de Machine Learning, los factores que inciden en la citación de los artículos de las revistas biomédicas colombianas indexadas en Scopus. En la revisión narrativa realizada para la construcción del marco teórico de este trabajo no se hallaron trabajos similares cuyo objeto de estudio fueran las publicaciones biomédicas colombianas. La búsqueda se realizó en Scopus, Pubmed vía Medline y Google Scholar, en español y en inglés, con las palabras clave "citación", "Machine Learning" y "Colombia".

## Metodología

Empleamos la metodología CRISP-DM, la cual está estandarizada y es de común aplicación en proyectos de analítica de datos (Martínez-Plumed *et al*., 2021). Sus etapas son: comprensión del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Para el análisis descriptivo de

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

citas recurrimos al programa *Publish or Perish* (Harzing, 2007), para el análisis descriptivo de otras variables, limpieza de datos, construcción y evaluación de modelos utilizamos *Python 3*.

### *Entendimiento de los datos*

Utilizamos el conjunto de datos "Revistas Indexadas, Índice Nacional Publindex 2017 – 2022" disponible en el portal de Datos Abiertos del Estado colombiano, el cual tiene información sobre las revistas colombianas especializadas (Datos Abiertos Colombia, 2022). También usamos la API (Application Programming Interfaces) de la base de datos Scopus para obtener información de los artículos en las revistas utilizando la librería 'pybliometrics' (Rose y Kitchin, 2019).

El conjunto de datos de Publindex cuenta con un total de 6276 registros y 34 variables. Al filtrar las revistas indexadas en 2022 y que pertenecen a la gran área 'Ciencias Médicas y de la Salud' se obtiene un total de 33 registros; estos fueron contrastados con el conjunto de datos de Scopus, el cual contaba con 18338 revistas indexadas hasta el 31 de mayo de 2023. Cruzamos esta información con la proveniente de Publindex a través de los ISSN obteniendo así un total de 20 revistas biomédicas colombianas.

De estas 20 revistas se obtuvo información de los artículos publicados entre el 1 de enero de 2019 y el 31 de mayo de 2023; rango máximo de tiempo en el que las 20 revistas tenían presencia en Scopus. Las variables obtenidas fueron identificadores del artículo (DOI, EID, Pubmed ID), de contenido (título, resumen, palabras clave, tipo de artículo, número de páginas, financiación), de los autores (nombres, afiliación, país, número de autores, índices *h* del primer y del último autor), de la publicación (fecha de publicación, número, volumen), el número de citaciones a la fecha y sobre el acceso al artículo (si es Open Access y qué tipo de acceso). Finalmente, se elaboró un análisis descriptivo de los datos y una matriz de correlaciones.

### *Preparación de los datos*

Retiramos las variables de identificación del artículo, país, número, volumen, título y resumen y se crearon dos nuevas variables binarias: una para establecer en cuáles artículos existía colaboración internacional y otra para identificar en los que participaba más de una institución. Categorizamos la variable de número de citas como binaria (tiene citas o no) y la variable de afiliaciones fue reemplazada por dos variables indicando la afiliación del primer y del último autor respectivamente.

Las palabras clave se procesaron normalizándolas (todas en minúscula) y vectorizándolas (codificar palabras como números) utilizando el vectorizador

TF-IDF de Scikit-Learn (Pedregosa *et al*., 2011); empleamos este instrumento ya que estima la relevancia que las palabras pueden tener en un conjunto de documentos. Se identificaron datos extremos en variables numéricas (índice *h* del primer autor, índice *h* del último autor, número de páginas y número de autores), tales se escalaron con RobustScaler de Scikit-Learn (Pedregosa *et al*., 2011).

Las variables categóricas nominales (tipo de artículo, nombre de la revista, tipo de Open Access, afiliaciones de primer y último autor) fueron transformadas en variables indicadoras con 'pandas.get_dummies' (The pandas development team, 2023).

## *Modelado*

Dividimos los datos en conjunto de entrenamiento y de prueba en una relación de 80/20. Empezamos probando dos modelos, uno que incluía todas las variables (Modelo 1) y uno sin palabras clave (Modelo 2). Para seleccionar el algoritmo adecuado para clasificar los datos, realizamos una validación cruzada para cada modelo utilizando la clase 'Classification' de *PyCaret* (Moez, 2020); esta estrategia de selección de algoritmos permite estimar el desempeño del modelo aplicado a otros datos diferentes a los del conjunto de entrenamiento. Empleamos la función 'compare_models' de la clase mencionada por su facilidad de aplicación e interpretación (*Figura 1*).

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| **gbc** | Gradient Boosting Classifier | 0.7989 | 0.8898 | 0.8084 | 0.7112 | 0.7557 | 0.5862 | 0.5908 | 10.8640 |
| **lightgbm** | Light Gradient Boosting Machine | 0.7928 | 0.8846 | 0.7633 | 0.7191 | 0.7387 | 0.5676 | 0.5702 | 2.3910 |
| **ada** | Ada Boost Classifier | 0.7885 | 0.8769 | 0.7757 | 0.7055 | 0.7372 | 0.5612 | 0.5649 | 3.3900 |
| **xgboost** | Extreme Gradient Boosting | 0.7855 | 0.8767 | 0.7666 | 0.7034 | 0.7331 | 0.5545 | 0.5565 | 5.5840 |
| **rf** | Random Forest Classifier | 0.7846 | 0.8727 | 0.6552 | 0.7528 | 0.6999 | 0.5333 | 0.5370 | 4.3430 |
| **dt** | Decision Tree Classifier | 0.7625 | 0.7526 | 0.7103 | 0.6868 | 0.6968 | 0.5020 | 0.5037 | 1.1890 |
| **et** | Extra Trees Classifier | 0.7137 | 0.7646 | 0.4240 | 0.7144 | 0.5312 | 0.3442 | 0.3689 | 5.9840 |
| **ridge** | Ridge Classifier | 0.7068 | 0.0000 | 0.5682 | 0.6339 | 0.5979 | 0.3684 | 0.3708 | 1.1500 |
| **knn** | K Neighbors Classifier | 0.6864 | 0.7225 | 0.5658 | 0.5961 | 0.5796 | 0.3300 | 0.3310 | 1.1640 |
| **lr** | Logistic Regression | 0.6743 | 0.7109 | 0.4261 | 0.6082 | 0.5004 | 0.2701 | 0.2797 | 19.1320 |
| **dummy** | Dummy Classifier | 0.6163 | 0.5000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.7020 |
| **lda** | Linear Discriminant Analysis | 0.5614 | 0.5525 | 0.5198 | 0.4420 | 0.4766 | 0.1044 | 0.1054 | 10.9630 |
| **svm** | SVM - Linear Kernel | 0.5368 | 0.0000 | 0.3899 | 0.1592 | 0.2255 | 0.0149 | 0.0169 | 2.3900 |
| **nb** | Naive Bayes | 0.5320 | 0.5461 | 0.6076 | 0.4233 | 0.4985 | 0.0854 | 0.0908 | 0.9360 |
| **qda** | Quadratic Discriminant Analysis | 0.4265 | 0.5076 | 0.8539 | 0.3998 | 0.5116 | 0.0165 | 0.0178 | 5.5650 |

*Figura 1.* Ejemplo de validación cruzada con *PyCaret* (Modelo 1)
Fuente: elaboración de los autores

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

Seleccionado el algoritmo, se realizó entrenamiento y ajuste. Ya que es esperado que el tiempo guarde relación con el número de citas de un artículo (Aksnes, Langfeldt y Wouters 2019), evaluamos modelos individuales por año.

### *Evaluación*

Revisamos precisión, retorno, F1 score (evaluación predictiva menos sesgada que la precisión), curvas ROC (representación gráfica de la proporción de verdaderos positivos que permite evaluar la capacidad de predicción de los modelos) y área bajo la curva (AUC, medición del acierto en la predicción del evento). Las importancias de las características empleadas para la clasificación se presentan como la contribución relativa de cada característica a la clasificación del modelo, con valores más altos representando una mayor importancia.

## Presentación y análisis de resultados

Fueron publicados 4 904 artículos en el periodo de análisis de estas revistas. Encontramos 5 140 citaciones, con 1 285 citaciones promedio por año y 1,05 citaciones por artículo; en promedio, el número de autores por artículo fue 4,35. Se eliminaron registros con información faltante y analizamos 4 130 artículos (*Tabla 1*).

| Variable | Artículos no citados | Artículos citados |
|---|---|---|
| **Número de autores (author_count) Mediana (IQR*)** | 4 (3) | 4 (3) |
| **Número de páginas (num_pag) Mediana (IQR)** | 5(7) | 7 (6) |
| **Índice *h* del primer autor (h_index_first) Mediana (IQR)** | 1 (3) | 2 (4) |
| **Índice *h* del último autor (h_index_last) Mediana (IQR)** | 2 (6) | 4 (9) |
| **Con financiación (fund_int) n (%)** | 260 (54,3) | 218 (45,6) |
| **Tipo de artículo (subtypeDescription) n (%)** | | |
| **Artículo original** | 1 910 (61,6) | 1 188 (38,3) |
| **Reporte/Series de caso** | 353 (73,2) | 129(26,7) |
| **Revisión** | 235 (50,1) | 234 (49,8) |
| **Editorial** | 17 (56,6) | 13 (43,3) |
| **Guías** | 11 (68,7) | 5 (31,2) |

| | | |
|---|---|---|
| **Otros** | 22 (62,9) | 13 (37,1) |
| | | |
| **Con acceso abierto (openaccess) n (%)** | 1 983 (61,1) | 1 260 (38,8) |
| **Tipo de acceso abierto (freetoreadLabel) n (%)** | | |
| **Bronce** | 182 (61,9) | 112 (38,1) |
| **Dorado** | 1 028 (72,8) | 383 (27,1) |
| **Verde** | 741 (49,2) | 764 (50,7) |
| **Dorado híbrido** | 23 (71,8) | 9 (28,1) |
| **Sin dato** | 574 (61,6) | 314 (35,3) |
| **Colaboración (colab_inst; colab_inter) n (%)** | | |
| **Con colaboración entre instituciones** | 1 643 (60,3) | 1 081 (39,6) |
| **Con colaboración internacional** | 325 (54,8) | 268 (45,1) |

*IQR: rango intercuartílico

*Tabla 1.* Distribución de las variables según proporción de artículos citados
Fuente: elaboración de los autores

Encontramos más artículos con citaciones en aquellos con más páginas, en revisiones, en los contenidos en revistas con índices más altos y en los accesos 'verdes'.

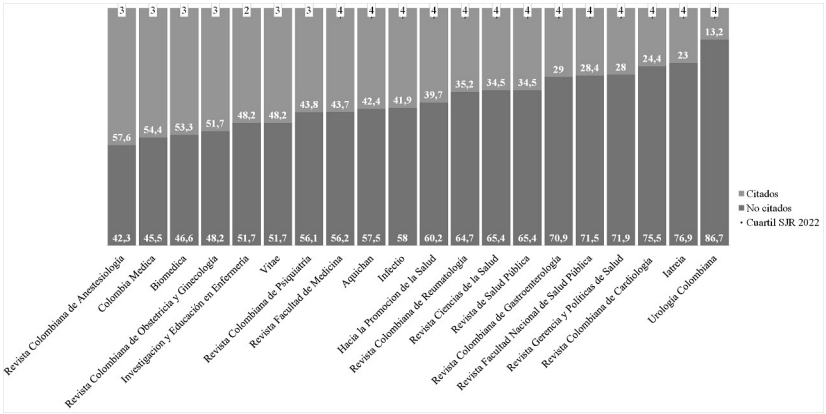


*Figura 2.* Proporción de artículos citados por revista
Fuente: elaboración de los autores

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

Por publicación, la mayor citación está en revistas de los cuartiles 2 y 3 de Scimago Journal & Country Rank (SJR) *(Figura 2*). La correlación más significativa puede verse entre el año de publicación y la variable de citación. También se observan otras correlaciones ligeramente significativas entre el índice *h* del primer autor y del último autor y entre colaboración internacional y colaboración interinstitucional (*Figura 3*).

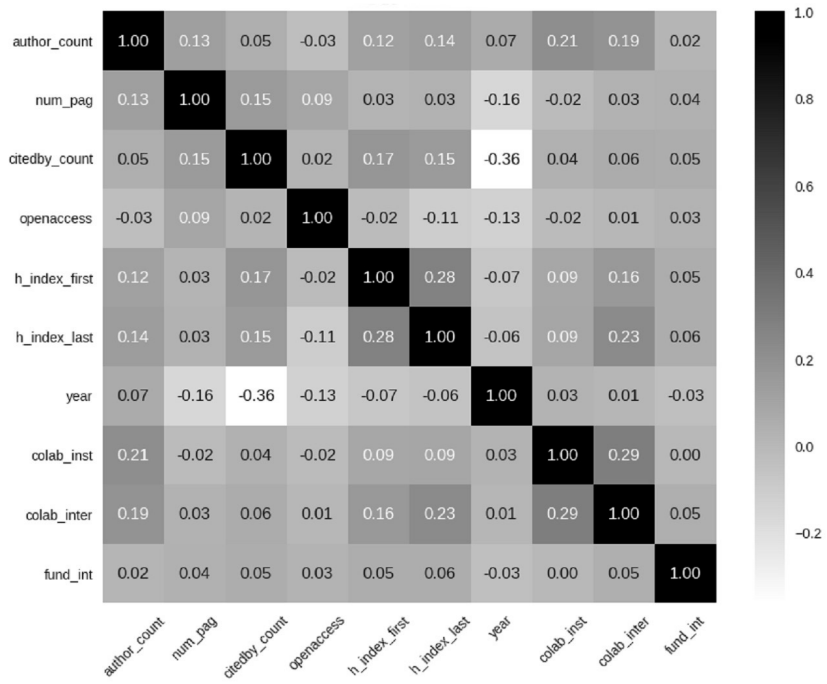


*Figura 3*. Matriz de correlación
Fuente: elaboración de los autores

Para los Modelos 1 y 2, el algoritmo seleccionado fue 'Gradient Boosting Classifier' (Pedregosa *et al*., 2011) (*Tabla 2*).

INVESTIGACIÓN BIBLIOTECOLÓGICA, vol. 38, núm. 99, abril/junio, 2024, México, ISSN: 2448-8321, pp. 89-107

| Modelo | Descripción | Algoritmo | Exactitud | F1 score | | Precisión | | Recall | | AUC | Muestra |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| 1 | Todas las variables | Gradient Boosting Classifier | 0,79 | 0,82 | 0,74 | 0,87 | 0,69 | 0,78 | 0,8 | 0,8821 | 4 130 |
| 2 | Se retiran palabras clave | | 0,78 | 0,82 | 0,73 | 0,86 | 0,69 | 0,78 | 0,79 | 0,8809 | 4 130 |
| 1a | Todas las variables, 2019 | | 0,74 | 0,65 | 0,79 | 0,86 | 0,69 | 0,52 | 0,93 | 0,7982 | 872 |
| 1b | Todas las variables, 2020 | | 0,77 | 0,68 | 0,82 | 0,96 | 0,71 | 0,52 | 0,98 | 0,8427 | 1 038 |
| 1c | Todas las variables, 2021 | | 0,71 | 0,78 | 0,6 | 0,8 | 0,56 | 0,75 | 0,63 | 0,8055 | 939 |
| 1d | Todas las variables, 2022 | Light Gradient Boosting Machine | 0,78 | 0,87 | 0,32 | 0,86 | 0,33 | 0,87 | 0,3 | 0,8183 | 959 |
| 1e | Todas las variables, 2023 | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 322 |

*Tabla 2.* Desempeño de los modelos
Fuente: elaboración de los autores

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

El desempeño del modelo disminuyó ligeramente al retirar las palabras clave; los índices *h* de los autores y el año son las características de mayor importancia en los modelos (*Figura 4*).

En los Modelos 1a, 1b y 1c empleamos 'Gradient Boosting Classifier', para los Modelos 1d y 1e usamos 'Light Gradient Boosting Machine'. El modelo 1e presentó un sobreajuste debido al reducido tamaño de muestra. El desempeño de cada modelo fue variable, obteniendo la mejor clasificación en el Modelo 1b (*Tabla 2*). Las curvas ROC se complementaron con los datos de AUC y hallamos que el mejor desempeño estaba en el Modelo 1, posiblemente por tener el tamaño de muestra más grande y el uso de todas las variables disponibles. En todos los casos, las curvas ROC muestran una capacidad de clasificación mayor a la esperada por azar (*Figura 5*).

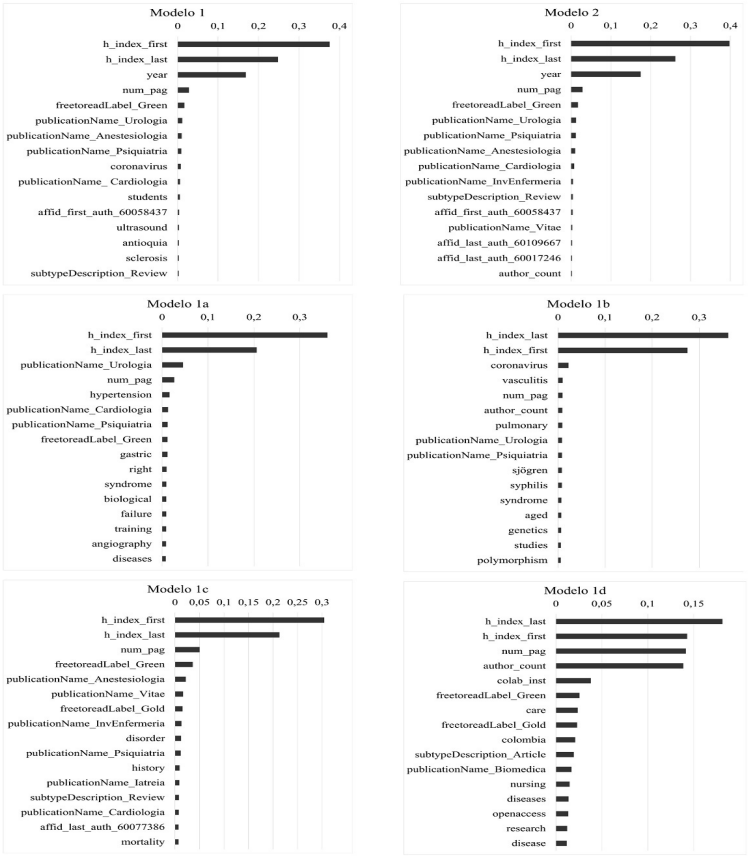


*Figura 4.* Principales características de importancia en cada modelo
Fuente: elaboración de los autores

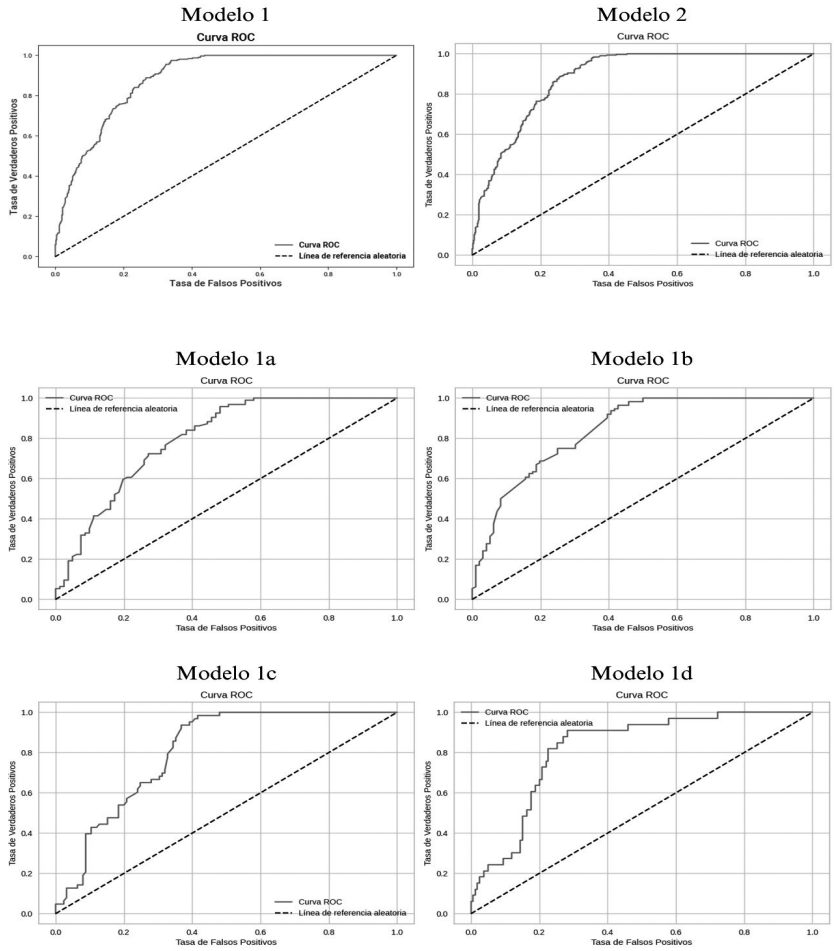


*Figura 5.* Curvas ROC de los modelos
Fuente: elaboración de los autores

Los índices *h*, el número de páginas, que el artículo tenga acceso 'verde' y el número de autores son factores relevantes, adicionalmente, y dependiendo del año, el hecho de pertenecer a ciertas publicaciones también fue una característica importante (*Figura 4*). Se resalta la pertenencia a la *Revista Colombiana de Psiquiatría* como característica de importancia en todos los modelos. Las palabras clave de importancia para cada modelo varían según el año. Colaboración internacional o institucional y financiación solo mostraron importancia en el Modelo 1d.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

## Discusión

En todos los modelos fueron características de importancia los índices *h* del primer y el último autor, este hallazgo coincide con el de Grover, Raman y Stubblefield (2014), quienes encontraron que el reconocimiento del autor fue el predictor más significativo de citación (1448). Fu y Aliferis (2010) hicieron una labor similar en literatura biomédica incluyendo al último y al primer autor y sus afiliaciones como características en modelos de Machine Learning, siendo las citaciones del último autor de gran importancia (264). A diferencia de estos autores, en los modelos propuestos no se encontró que la inclusión de las afiliaciones fuera relevante. El índice *h* es considerado como un indicador de la reputación del autor y del impacto de su producción académica previa (Cronin y Meho, 2006) y se cree que ese reconocimiento atrae nuevas citaciones por el efecto Mateo (Merton, 1988) o porque su experiencia le facilita realizar estudios relevantes de calidad y divulgarlos adecuadamente. Sin embargo, Grover y su equipo (2014) señalan que, bajo este planteamiento, algunos trabajos relevantes podrían pasar desapercibidos si no tienen un autor reconocido (1450).

Los accesos de tipo 'verde' y 'dorado' tuvieron cierta relevancia en todos los modelos excepto en el 1b. Hay duda sobre si los artículos de acceso abierto tienden a ser más citados (conocido como sesgo FUTON), pues la mayoría de los estudios al respecto no suelen ser comparables entre sí y sus resultados no son generalizables (Langham-Putrow, Bakker y Riegelman, 2021), empero, diferentes autores han concluido que posiblemente existe este sesgo, pero tal es variable y depende del tipo de acceso abierto y área temática (Basson, Blanckenberg y Prozesky, 2021). Piwowar y su equipo (2020) hallaron que los artículos con acceso 'verde' o 'híbrido' reciben citaciones hasta 30% por encima del promedio de citaciones relativas al compararlo con otros tipos de acceso, no obstante, con el paso del tiempo, el promedio de citaciones relativas tiende a disminuir en los artículos con acceso 'dorado' y se mantiene estable en los de 'verde' (14). En el caso del presente estudio encontramos una relación entre el acceso 'verde' y la presencia de citaciones en todos los modelos excepto en el 1b.

El año 2020 es atípico en la literatura biomédica, pues se dio acceso abierto a todas las publicaciones relacionadas con coronavirus y se priorizó la publicación de artículos relacionados con el tema para brindar ayuda en los momentos más críticos de la pandemia (Arrizabalaga *et al.*, 2020). El Modelo 1b muestra el mejor desempeño, posiblemente porque cuenta con una muestra más grande, así como con un mayor número de citaciones (556 artículos citados) que permite un mejor entrenamiento. Otra particularidad es que prácticamente todas las variables de importancia son palabras clave. Al retirar las palabras clave el desempeño del Modelo 2 disminuyó ligeramente; esto podría indicar que los temas presentes

en los artículos se relacionan con la citación. Es más evidente cuando se observa que la palabra "coronavirus" adquiere mayor relevancia en el modelo para el año 2020 y en otros años las palabras presentes son diversas. Fu y Aliferis (2010) incluyeron términos MeSH en su trabajo obteniendo resultados similares; los términos resultaban de relevancia para predecir citación y eran pocos aquellos que se repetían en todos los modelos (265).

Los primeros dos modelos dieron gran importancia al año y la matriz de correlación mostró una alta correlación negativa del año con la variable objetivo; tal parece indicar que mientras más antiguo sea el artículo se espera que tenga más citaciones, pues ha tenido una mayor oportunidad de ser leído y acumular citaciones en comparación con un artículo publicado recientemente. Este es un hallazgo recurrente en análisis bibliométricos y da cuenta de la importancia de considerar el marco temporal al analizar citaciones (Aksnes, Langfeldt y Wouters 2019).

El número de autores fue relevante para los Modelos 1b y 1d, lo cual coincide con hallazgos en otros estudios (Figg *et al*., 2006; Bordons, Aparicio y Costas, 2013); un número más grande de autores puede atraer más citaciones debido a que el efecto Mateo es mayor, además, si un artículo cuenta con un mayor número de autores debería tener más complejidad e incluso calidad al contar con más apoyo en su desarrollo. El número de autores no resulta relevante en todos los análisis; en una revisión se encontró que, aunque el número de autores puede tener un impacto, este puede no ser significativo en todas las áreas temáticas (Onodera y Yoshikane, 2015). En otros contextos esta variable se ha utilizado como indicador de colaboración nacional e internacional; en nuestros modelos las variables para evaluar estos aspectos no fueron de gran relevancia. Como sucede en nuestro caso, He (2009) descartó que la colaboración internacional tuviera algún impacto sobre la citación demostrando a través de modelos de regresión que resulta igual a la colaboración nacional (2162).

En el estudio de Grover y su equipo (2014), el número de páginas fue incluido como una variable que indicaba el nivel con el que el autor lograba aclarar las ideas presentadas en el trabajo y descubrieron que los artículos más largos tienden a tener un mayor número de citaciones (1450). En concordancia con estos hallazgos, el número de páginas fue otra característica de gran importancia en todos los modelos; este suele ser más alto en los artículos de revisión y las guías de práctica clínica, por lo que la importancia de esta característica podría estar relacionada con el hecho de representar a estos artículos altamente citados, sin embargo, dentro de los modelos revisados no encontramos que el tipo de artículo tuviera relevancia. El número de páginas podría representar también un artículo de mayor complejidad con información relevante que atrae más citaciones. Cabe aclarar que usualmente las revistas regulan el número de páginas y palabras que contiene un artículo, un tope que es mayor en el caso de las revisiones y las

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

guías de práctica clínica. Sugerimos profundizar en este aspecto en futuras investigaciones.

El pertenecer a una revista no parece tener gran importancia en los modelos y cuando aparecen nombres de revistas encontramos que, en el año evaluado, las revistas tuvieron un alto número de citaciones o publicaron artículos relacionados con las palabras clave de alta relevancia. Por ejemplo, la *Revista Colombiana de Cardiología* publicó un mayor número de artículos relacionados con el tema "hipertensión" en 2019 (38,8% del total de artículos). La importancia del nombre de la publicación depende del factor de impacto que tenga (Onodera y Yoshikane, 2015). Para los modelos de este trabajo, la *Revista Colombiana de Psiquiatría* aparece con diferentes grados de significancia en cada uno y es la segunda revista con SJR más alto (0,358 en 2022). La revista con el SJR más alto, *Investigación y Educación en Enfermería*, solo se presenta con algo de importancia en los Modelos 2 y 1c; se plantea la posibilidad de que la diferencia de importancia entre estas dos revistas radique en las temáticas que publicaron en los años analizados.

Dentro de las limitaciones en el desarrollo de este trabajo está la imposibilidad de comprobar la veracidad de los metadatos de los artículos, pues no existe una manera de verificar si todos fueron cargados y recuperados correctamente. Adicionalmente, existen ciertas peculiaridades de la indexación de las revistas las cuales pudieron haber limitado la obtención de información, por ejemplo, los cambios de nombre o de ISSN no permiten tener certeza de su veracidad. Una última limitación que debe considerarse es la aplicabilidad de los resultados obtenidos, este análisis contempla únicamente revistas científicas biomédicas colombianas indexadas en Scopus y en Publindex, por lo tanto, la comparación de los resultados obtenidos con los de otras áreas o revistas debe realizarse con precaución.

Con los resultados observados puede concluirse que el factor que más incide en la citación de un artículo biomédico colombiano es la reputación de sus autores, lo cual refleja un paradigma presente en las publicaciones científicas: los autores más prestigiosos buscan publicar artículos que puedan tener gran visibilidad en las revistas más prestigiosas y esto trae, a su vez, mayor visibilidad y prestigio tanto a los autores como a las revistas. Aunque el prestigio es un factor importante se recomienda que no sea el único que motive a la lectura o a la aceptación de un artículo, pues tal podría limitar el crecimiento y la diversidad del área temática. El acceso abierto de tipo 'verde' y 'dorado' se presentaron como características con cierta importancia, posiblemente porque permiten una mayor visibilidad de los artículos.

En Latinoamérica, el acceso abierto tiene una gran relevancia, pues es considerado como una estrategia de divulgación científica y, al haber comprobado cierto nivel de utilidad para la obtención de citaciones, vale la pena considerarlo no solo como una recomendación para mejorar las métricas, sino también como

una herramienta de desarrollo científico. El número de páginas es la tercera variable que podría emplearse para mejorar el conteo de citaciones, aunque se recomienda contemplarla bajo las conclusiones expresadas por Grover y su equipo (2014): "…los autores usan el espacio que la revista deja a su disposición para explicar efectivamente ideas complejas e interesantes" (1450). Es decir, no se trata de alargar el contenido de un artículo sin un objetivo, sino de garantizar que las ideas principales sean explicadas con claridad. Finalmente, el análisis evidenció la relevancia de las palabras clave, especialmente en el Modelo 1b. Con base en este hallazgo, recomendamos estar al tanto de los temas más relevantes, ya que estos no solo tendrán una mejor visibilidad, sino que son de utilidad para las discusiones en torno al tema.

## Conclusiones

El prestigio y experiencia de los autores, representados en su índice *h*, se ven resaltados en este estudio como factores importantes que influyen en la citación; asimismo, el número de páginas y de autores también actúa como un posible indicador de la complejidad del artículo que, a su vez, repercute en los índices de citación. Reconocemos, además, la importancia de las temáticas del artículo, representadas por las palabras clave, como factor que motiva a su citación y también al papel que el acceso abierto ejerce como oportunidad para difundir el artículo y permitirle ser citado. Se han identificado unos factores comunes a los modelos de Machine Learning empleados, los cuales pueden considerarse como los más influyentes en la citación de un artículo biomédico colombiano.

Las variables presentadas como de alta importancia pueden tomarse en cuenta al momento de elaborar o publicar un artículo, sin embargo, no debe abandonarse la idea de citar artículos de calidad y contenido relevantes, independientemente de la reputación de los autores o la revista donde estén publicados y, también, no debe dejarse de fomentar esta práctica entre la comunidad académica en general.

## Referencias

Aksnes, Dag, Liv Langfeldt y Paul Wouters. 2019. "Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories". *SAGE Open* 9 (1): 1-17. https://doi.org/10.1177/2158244019829575

Alohali, Yousef, Mahmoud Samir Fayed, Tamer Mesallam, Yassin Abdelsamad, Fida Almuhawas y Abdulrahman Hagr. 2022. "A Machine Learning Model to Predict Citation Counts of Scientific Papers in Otology Field". *BioMed Research International* 2022: 1-12. https://doi.org/10.1155/2022/2239152

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

Anderson, Caleb, Kenneth Nugent y Christopher Peterson. 2021. "Academic Journal Retractions and the COVID-19 Pandemic". *Journal of Primary Care & Community Health* 12: 1-6
https://doi.org/10.1177/21501327211015592

Aphinyanaphongs, Yindalon, Alexander Statnikov y Constantin Aliferis. 2006. "A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents". *Journal of the American Medical Informatics Association* 13 (4): 446-55.
https://doi.org/10.1197/jamia.M2031

Arrizabalaga, Olatz, David Otaegui, Itziar Vergara, Julio Arrizabalaga y Eva Méndez. 2020. "Open Access of COVID-19-Related Publications in the First Quarter of 2020: A Preliminary Study Based in PubMed". *F1000Research* 9 (649): 1-34.
https://doi.org/10.12688/f1000research.24136.2

Basson, Isabel, Jaco Blanckenberg y Heidi Prozesky. 2021. "Do Open Access Journal Articles Experience a Citation Advantage? Results and Methodological Reflections of an Application of Multiple Measures to an Analysis by WoS Subject Areas". *Scientometrics* 126 (1): 459-84.
https://doi.org/10.1007/s11192-020-03734-9

Bordons, María, Javier Aparicio y Rodrigo Costas. 2013. "Heterogeneity of Collaboration and Its Relationship with Research Impact in a Biomedical Field." *Scientometrics* 96 (2): 443-66.
https://doi.org/10.1007/s11192-012-0890-7

Cáceres Castellanos, Gustavo. 2014. "La Importancia de publicar los resultados de investigación". *Revista Facultad de Ingeniería* 23 (37): 7-8.
https://www.redalyc.org/articulo.oa?id=413937008001

Crespo, Juan, Yungrong Li y Javier Ruiz-Castillo. 2012. "Differences in Citation Impact across Scientific Fields". *Working Papers Economic Series* 12 (6): 1-32.
https://e-archivo.uc3m.es/bitstream/handle/10016/14771/we1206.pdf?sequence=1

Crespo, Juan, Yungrong Li y Javier Ruiz-Castillo. 2013. "The Measurement of the Effect on Citation Inequality of Differences in Citation Practices across Scientific Fields". *PLOS ONE* 8 (3): 1-9.
https://doi.org/10.1371/journal.pone.0058727

Cronin, Blaise, y Lokman Meho. 2006. "Using the H-index to Rank Influential Information Scientists". *Journal of the American Society for Information Science and Technology* 57 (9): 1275-78.
https://doi.org/10.1002/asi.20354

Datos Abiertos Colombia. 2022. "Revistas Indexadas, Índice Nacional Publindex 2017 - 2022". Ciencia, Tecnología e Innovación. 28 de noviembre de 2022.
https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Revistas-Indexadas-ndice-Nacional-Publindex-2017-2/fsjb-9cah

Figg, William, Lara Dunn, David Liewehr, Seth Steinberg, Paul Thurman, Carl Barrett y Julian Birkinshaw. 2006. "Scientific Collaboration Results in Higher Citation Rates of Published Articles". *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 26 (6): 759-67.
https://doi.org/10.1592/phco.26.6.759

Fu, Lawrence, y Constantin Aliferis. 2010. "Using Content-Based and Bibliometric Features for Machine Learning Models to Predict Citation Counts in the Biomedical Literature". *Scientometrics* 85 (1): 257-70. https://doi.org/10.1007/s11192-010-0160-5

Grover, Varun, Roopa Raman y Adam Stubblefield. 2014. "What Affects Citation Counts in MIS Research Articles? An Empirical Investigation". *Communications of the Association for Information Systems* 34: 1435-56. https://doi.org/10.17705/1CAIS.03474

Harzing, Anne-Wil. 2007. *Publish or Perish*. V. 8. Windows. https://harzing.com/resources/publish-or-perish

He, Zi-Lin. 2009. "International Collaboration Does Not Have Greater Epistemic Authority". *Journal of the American Society for Information Science and Technology* 60 (10): 2151-64. https://doi.org/10.1002/asi.21150

Iqbal, Sehrish, Saeed-Ul Hassan, Naif Radi Aljohani, Salem Alelyani, Raheel Nawaz y Lutz Bornmann. 2021. "A Decade of In-Text Citation Analysis Based on Natural Language Processing and Machine Learning Techniques: An Overview of Empirical Studies". *Scientometrics* 126: 6551-99. https://doi.org/10.1007/s11192-021-04055-1

Langham-Putrow, Allison, Caitlin Bakker y Amy Riegelman. 2021. "Is the Open Access Citation Advantage Real? A Systematic Review of the Citation of Open Access and Subscription-Based Articles". *PLOS ONE* 16 (6): 1-20. https://doi.org/10.1371/journal.pone.0253129

Martínez-Plumed, Fernando, Lidia Contreras-Ochando, Cesar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana y Peter Flach. 2021. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories". *IEEE Transactions on Knowledge and Data Engineering* 33 (8): 3048-61. https://doi.org/10.1109/TKDE.2019.2962680

Martinovich, Viviana. 2020. "Indicadores de citación y relevancia científica: genealogía de una representación". *Dados. Revista de Ciências Sociais* 63 (2): 2-29. https://doi.org/10.1590/001152582020218

Merton, Robert King. 1988. "The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property". *Isis* 79 (4): 606-23. https://www.jstor.org/stable/234750

Mingers, John, y Loet Leydesdorff. 2015. "A Review of Theory and Practice in Scientometrics." *European Journal of Operational Research* 246 (1): 1-19. https://doi.org/10.1016/j.ejor.2015.04.002

Moez, Ali. 2020. *PyCaret: An Open Source, Low-Code Machine Learning Library in Python*. V. 1.0.0. https://www.pycaret.org

Navarrete, Luz, y Claudia Pérez. 2019. "Revistas biomédicas: desarrollo y evolución". *Revista Médica Clínica Las Condes* 30 (3): 219-25. https://doi.org/10.1016/j.rmclc.2019.04.002

Onodera, Natsuo, y Fuyuki Yoshikane. 2015. "Factors Affecting Citation Rates of Research Articles". *Journal of the Association for Information Science and Technology* 66 (4): 739-64. https://doi.org/10.1002/asi.23209

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss y Vincent Dubourg. 2011. "Scikit-Learn: Machine Learning in Python". *The Journal of Machine Learning Research* 12: 2825-30. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West y Stefanie Haustein. 2020. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles". *PeerJ* 6: 1-23. https://doi.org/10.7717/peerj.4375

Pradhan, Dinesh, Joyita Chakraborty y Subrata Nandi. 2019. "Applications of Machine Learning in Analysis of Citation Network". En *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data,* 330-33. https://doi.org/10.1145/3297001.3297053

Repiso, Rafael, Alicia Moreno-Delgado e Ignacio Aguaded. 2021. "Factors Affecting the Frequency of Citation of an Article". *Iberoamerican Journal of Science Measurement and Communication* 1 (1): 1-6. https://doi.org/10.47909/ijsmc.08

Ronda-Pupo, Guillermo Armando, Nelson Fernández-Vergara, Rodrigo Alda-Varas, Fernando Aurelio Álvarez-Castillo, Carlos Molina y Walter Sergio Terrazas-Núñes. 2022. "Evaluación del desempeño investigativo del Sistema Universitario Chileno 2006-2020". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 36 (91): 109-23. https://doi.org/10.22201/iibi.24488321xe.2022.91.58505

Rose, Michael, y John Kitchin. 2019. "Pybliometrics: Scriptable Bibliometrics Using a Python Interface to Scopus". *SoftwareX* 10: 100263. https://doi.org/10.1016/j.softx.2019.100263

Stephan, Paula, Reinhilde Veugelers y Jian Wang. 2017. "Reviewers Are Blinkered by Bibliometrics". *Nature* 544: 411-12. https://doi.org/10.1038/544411a

Su, Zhongqi. 2020. "Prediction of Future Citation Count with Machine Learning and Neural Network". En *2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 101-4. IEEE. https://doi.org/10.1109/IPEC49694.2020.9114959

The pandas development team. 2023. "pandas-dev/pandas: Pandas (v2.1.1)". *Zenodo*, 20 de septiembre de 2023. https://doi.org/10.5281/zenodo.8364959

*Para citar este texto:*

Sánchez-Bello, Nubia Fernanda, Jorge Enrique Mejía Quiroga y Constanza Beatriz Pérez-Martelo 2024. "Factores asociados a la citación de artículos biomédicos colombianos: análisis con machine learning". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 38 (99): 89-107. http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.99.58857