SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 3, (2025), pp. 415-425
Open Access. Online & Print

DOI: 10.5281/zenodo.11032825

PREDICTIVE MODELING OF JOURNAL EXCLUSION FROM SCOPUS BASED ON BIBLIOMETRIC AND EDITORIAL INDICATORS

Carla Isabel Lozano-Alvarado[1*], Dennis Alfredo Peralta-Gamboa[2]

[1]Universidad Estatal de Milagro, Milagro, Ecuador. E-mail: clozanoa@unemi.edu.ec, ORCID iD: https://orcid.org/0009-0003-1963-1339
[2]Universidad Estatal de Milagro, Milagro, Ecuador. E-mail: daperalt8@gmail.com, ORCID iD: https://orcid.org/0009-0009-0636-0094

Received: 22/11/2025
Accepted: 17/02/2026

Corresponding Author: Carla Isabel Lozano-Alvarado (clozanoa@unemi.edu.ec)

ABSTRACT

This study examines the editorial, bibliometric, and contextual factors associated with the exclusion of scientific journals from Scopus. Using a quantitative approach, a binary logistic regression model was applied to a balanced sample of 306 journals (153 active and 153 excluded journals). The analysis included variables such as the h-index, SC Imago Journal Rank (SJR), plagiarism detection policy, article processing charges (APC), and publisher's country. The results indicate that the h-index positively predicts retention in Scopus, while fee-waiver policies for authors from developing countries are associated with a lower probability of continuation. The model demonstrated a good predictive performance (AUC = 0.812), reinforcing its usefulness as a tool for editorial monitoring. These findings offer empirical evidence for strengthening sustainable editorial practices and improving scientific governance in international indexing systems.

KEYWORDS: Journal Exclusion, Scopus; Logistic Regression, Editorial Quality, Bibliometrics, Open Access.

1. INTRODUCTION

Indexing in scientific databases such as Scopus is a key indicator of quality, visibility, and academic impact (Mongeon & Paul-Hus, 2016). To remain on these platforms, journals must meet strict criteria including a robust peer-review process, regular publication frequency, editorial board diversity, and strong citation metric performance. Editorial ethics, transparency in publication costs, and compliance with international standards were also evaluated according to Elsevier's guidelines (2023).

In recent years, the number of journals excluded from these databases has increased significantly because of non-compliance with editorial, ethical, or bibliometric standards (Valz Gris et al., 2024; Pandita & Singh, 2023). Predatory journals that prioritize profit over academic integrity are a major factor in exclusion. These journals often engage in deceptive practices, such as false peer review claims and misleading editorial information, which leads to their removal from accredited databases (Soundarapandian, 2023; Severin & Low, 2019).

This phenomenon raises concerns within the academic community by questioning the fairness, consistency, and transparency of evaluation and retention processes. In a global context, where low-quality journals, such as predatory and hijacked journals, proliferate, platforms such as Scopus, Web of Science, and PubMed have strengthened their exclusion mechanisms to protect scientific integrity (Grudniewicz et al., 2019; Macháček & Srholec, 2022).

Understanding the factors behind journal exclusion is important for editors, managers, and policymakers. Exclusion not only affects a journal's visibility, but also impacts researchers' careers, institutional evaluations, and eligibility for funding, especially in the Global South, where structural limitations hinder compliance with international standards (Anderson et al., 2022; Williams et al., 2023; Quiroga-Garza et al., 2022).

This study aimed to identify the main bibliometric, editorial, and contextual factors that contribute to journal exclusion from Scopus. Using a quantitative and explanatory approach, a binary logistic regression model is applied to a balanced sample of active and excluded journals. The variables analyzed included the h-index, SJR, average time from submission to publication, plagiarism detection policies, article processing charges (APC), fee waivers for authors from developing countries, and the geographical location of the publisher.

It is hypothesized that factors such as a low h-index, lack of institutional financial support, and prolonged publication times significantly increase the likelihood of exclusion. Unlike studies that focus on initial inclusion, this research addresses the exclusion of previously indexed journals from a predictive and empirical perspective. The results provide tools for editorial monitoring and contribute to the design of more equitable, sustainable, and evidence-based science policies.

To guide the reader through the structure of this paper, the following sections are organized as follows: Section 1.1 presents a theoretical and literature review on journal exclusion and editorial indicators. Section 2 details the methodology, including the data collection, research design, and statistical modeling techniques. Section 3 reports the main results of the logistic regression analysis as well as complementary models and visualizations. Section 4 discusses the findings in light of the current scientific publishing practices and editorial policy implications. Finally, Section 5 presents the conclusions of the study and proposes directions for future research.

1.1. Theoretical and Literature Review

1.1.1. Bibliometric Indicators and Journal Quality

Bibliometric indicators, such as the h-index and SCImago Journal Rank (SJR), are widely used to assess journal quality and impact. The h-index, introduced by Hirsch (2005), combines productivity with citation impact and remains a key metric in editorial evaluation (Ruscio, 2016). Jamali et al. (2014) argue that SJR reflects journal prestige but is influenced by contextual factors such as region and language publication. However, some studies caution that these indicators, while useful, have limitations across disciplines and may not fully capture the editorial quality.

1.1.2. Editorial Practices and Publishing Ethics

Editorial integrity plays a critical role in maintaining indexing. Plagiarism detection, peer review rigor, and transparency in publication policies are essential elements of editorial governance (Soundarapandian, 2023; Wager, 2012). Guidelines from COPE and DOAJ encourage ethical standards, yet Marina and Sterligov (2021) observed that anti-plagiarism practices are uniformly applied in top-tier journals, limiting their discriminatory power. Open access policies, including Article Processing Charges (APCs) and fee waivers, also shape editorial decisions. Shen and Björk (2015) emphasize that while APCs support sustainability, over-reliance on them or extensive waivers may compromise their financial viability.

1.1.3. Structural Inequalities in Indexing Systems

Several studies have highlighted the structural barriers faced by journals from the Global South in meeting indexing standards. Tennant (2020) critiques the underrepresentation of non-Western and non-English journals in Scopus and the Web of Science. These platforms often overlook legitimate regional journals because of the systemic biases in language, geography, and funding. Anderson et al. (2022) and Quiroga-Garza et al. (2022) underscore the challenges in editorial professionalization and resource access. Despite efforts from initiatives such as SciELO and Plan S (cOAlition, 2018), inequities persist, affecting visibility, funding eligibility, and academic career progression.

2. METHODOLOGY

2.1 Data Collection and Preparation

The database was built in R (version 4.4.2) using the openxlsx and dplyr packages. The list of active journals in Scopus as of March 2025 was downloaded, including bibliometric data and using the journal title as the primary identifier. This information was cross-referenced with the historical SCImago Journal Rank (SJR) database from 1999 to 2024 via a left_join, classifying journals as active (present in SJR 2024) or inactive (last appearance in SJR 2023 or earlier), identifying 9,131 inactive journals.

To obtain a balanced sample, 9,131 active journals were randomly selected. This sample was then cross-referenced with the Directory of Open Access Journals (DOAJ), using the title as the key, to incorporate editorial data such as plagiarism detection policies, article processing charges (APCs), fee waivers, and the average time from submission to publication. After merging, complete information was obtained for 1,991 journals (1,838 active and 153 inactive). To balance the sample, 153 active journals were randomly selected, resulting in a final dataset of 306 journals (see Figure 1). Random undersampling was chosen for its simplicity and low risk of overfitting compared to techniques like SMOTE or inverse weighting, making it suitable for imbalanced class scenarios in bibliometric studies (Buda et al., 2018).

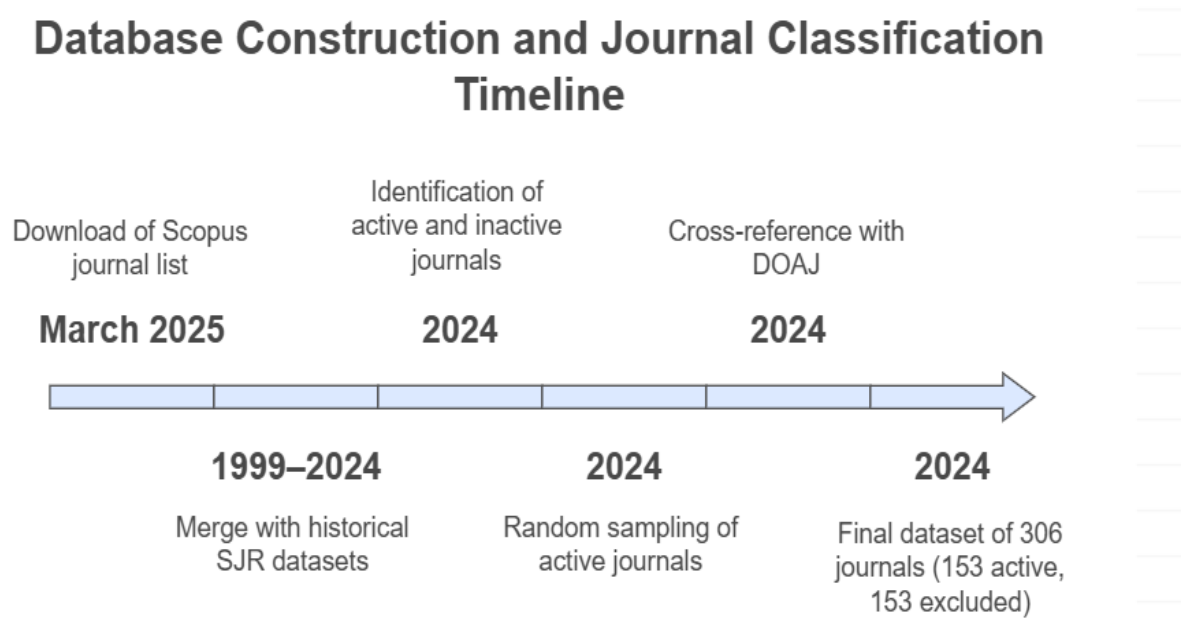

Figure 1: Diagram *of the Methodological Flow for Sample Construction and Selection.*

2.2. Research Design

This study adopts a quantitative, explanatory, and cross-sectional design to identify predictive factors for journal exclusion from Scopus. Causal relationships were modeled between independent variables (editorial and bibliometric) and a binary dependent variable (indexing status: 1 = active, 0 = excluded). Multicollinearity was checked using the Variance Inflation Factor (VIF), with values below 2 indicating low collinearity.

A binary logistic regression was performed using R's glm() function (binomial family). The independent variables are detailed in Table 1.

Table 1: Variables Included in the Logistic Regression Model and Their Description.

Variable	Type	Description
SJR (transformed)	Continuous	Citation-based prestige index
h-index (transformed)	Continuous	Cumulative productivity and citation impact
Publication weeks (transformed)	Continuous	Average time from submission to publication
Developing country	Binary	1 = country classified by World Bank
Plagiarism detection	Binary	1 = journal reports using anti-plagiarism tools
APC charges	Binary	1 = APCs are required
Fee waivers	Binary	1 = waivers offered to authors from developing countries

2.3. Cross-Validation and Complementary Approach

To ensure model robustness and prevent overfitting, stratified k-fold cross-validation (k = 10) was applied with an 80% training and 20% testing split, using the train () function from the caret package in R. The evaluation metrics (AUC, accuracy, Kappa) demonstrated stability, confirming the model's generalizability. A decision tree (CART) and a penalized LASSO regression were also employed to explore non-linear interactions and validate variable selection. The decision tree identified critical thresholds, such as h-index < 0.5 being associated with a higher risk of exclusion, while LASSO confirmed the h-index as the main predictor, discarding less relevant variables. Both approaches supported the findings of the logistic regression. The evaluation metrics included accuracy, sensitivity, specificity, Kappa coefficient, Nagelkerke's pseudo R^2, and AUC, following best practices for binary classification in imbalanced datasets (He & Garcia, 2009; Chawla et al., 2002).

3. RESULTS

3.1. Model Analysis

A binary logistic regression model was estimated to identify the factors associated with the exclusion of scientific journals from the Scopus index. The dependent variable was indexing status (1 = active; 0 = excluded), while the independent variables included bibliometric indicators, editorial characteristics, and contextual factors. To improve the linearity of relationships between continuous variables and the logit, square root transformations were applied to SJR, h-index, and weeks to publication.

Specifically, the transformation function used was $\sqrt{x}$, where x is the original value of the SJR and h-index. This choice helped reduce skewness and ensured better adherence to the linearity assumption of the logit model.

In addition, several predictors, such as plagiarism detection, APC charges, and publication time, remained in the model, despite their lack of statistical significance. This decision was grounded in both the theoretical relevance and practical considerations. These variables have been previously highlighted in the literature as meaningful editorial and bibliometric indicators (e.g., Shen & Björk, 2015; Wager, 2012), and their inclusion allowed for consistent comparison across complementary models (LASSO, decision tree) and facilitated the interpretability of visualizations such as the forest plot. The model was fitted to a balanced sample of 306 scientific journals (153 active and 153 excluded). The likelihood ratio test yielded a chi-square value of 102.87 ($p < 0.001$), indicating that the model with predictors fits significantly better than the null model. The AIC value was 337.33, and Nagelkerke's pseudo R^2 was 0.381, indicating moderate explanatory power.

3.2. Significant Predictors

Of the eight predictors included, **two were statistically significant (Table 2)**

- h-index (transformed)

1. Coefficient: $\beta = 0.505$
2. p-value < 0.001
3. OR = 1.657
4. **Interpretation** Each unit increase in the square root of the h-index increases the likelihood of a journal remaining indexed by 65.7%. This supports the role of the h-index as a robust indicator of editorial stability and prestige.

- Fee waivers for authors from developing countries

a. Coefficient: $\beta = -1.016$
b. p-value = 0.042
c. OR = 0.362
d. **Interpretation** The existence of such policies is associated with a 63.8% reduction in the likelihood of being retained. Although inclusive, these measures may reflect structural economic limitations that undermine editorial sustainability.

In addition, the variable developing countries showed a marginally significant trend:

- β = 0.597, p = 0.094, OR = 1.817

This positive, though inconclusive, association may indicate progressive improvements in editorial quality in Global South contexts. The variable "developing country" showed a marginally significant positive association with journal retention (β = 0.597, p = 0.094, OR = 1.817). While not statistically significant at the conventional 0.05 level, this trend may reflect gradual improvements in editorial quality among journals based in low and middle-income countries. It suggests potential progress in professionalization and compliance with indexing standards, although further research with larger samples is needed to confirm this effect.

Table 2: Estimated Coefficients of the Binary Logistic Regression Model.

Predictor Variable	β Coefficient		z-value	p-value	Exp(β)	Interpretation
Intercept	-2.334	0.706	-3.308	0.001	0.097	–
$\sqrt{SJR}$	0.745	0.565	1.318	0.188	2.106	Not significant
$\sqrt{\text{h} - \text{index}}$	0.505	0.066	7.641	<0.001	1.657	Increases probability of retention
Developing country (1 = yes)	0.597	0.356	1.676	0.094	1.817	Marginally positive association
Plagiarism detected (1 = yes)	0.219	0.347	0.631	0.528	1.245	Not significant
$\sqrt{\text{Weeks to publish}}$	0.064	0.124	0.517	0.605	1.066	Not significant
Fee waivers (1 = yes)	-1.016	0.500	-2.032	0.004	0.362	Decreases probability of retention

3.3. Model Validation

To evaluate the model's predictive performance, a stratified partition of the data was performed: 80% for training and 20% for testing. **On the test set, the model achieved**

- Overall accuracy: 71.7%
- Sensitivity: 76.7% (correct classification of active journals)
- Specificity: 66.7% (correct classification of excluded journals)
- Kappa coefficient: 0.43 (moderate agreement)
- Area Under the ROC Curve (AUC): 0.812 (see Figure 2).

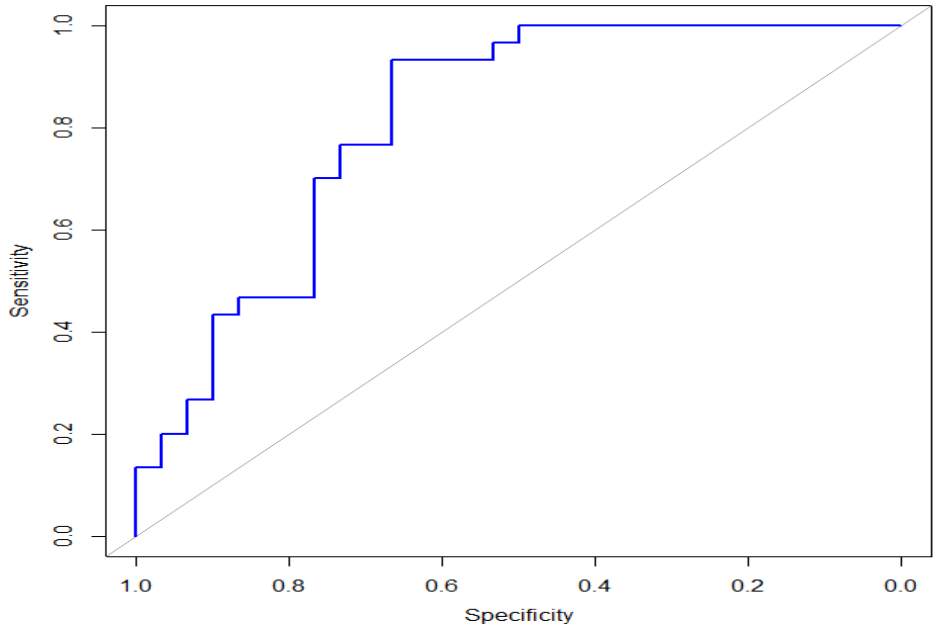

Figure 2: ROC Curve of the Binary Logistic Regression Model.

3.4. Complementary Visualization

To enhance the analytical interpretation and provide a comprehensive view of the findings, visual representations were incorporated to complement the inferential statistical approach. These visuals help triangulate the regression model results with descriptive patterns and facilitate intuitive identification of relevant relationships. Figure 3 presents a forest plot displaying the estimated coefficients from the logistic model along with their 95% confidence intervals. This tool clearly shows the magnitude and direction of each predictor's effect on

the probability of remaining indexed in Scopus as well as its statistical significance. The transformed h-index is the only predictor with a statistically significant positive coefficient and a confidence interval that does not cross zero, reinforcing its role as a consistent indicator of editorial stability. By contrast, the fee waiver variable shows a significant negative coefficient, while predictors such as SCImago Journal Rank (SJR), publication weeks, and plagiarism detection policies do not show significance. This type of visualization effectively synthesizes model results and allows for ranking predictors by statistical relevance, offering a clear and concise interpretation that is especially useful in multidisciplinary communication settings.

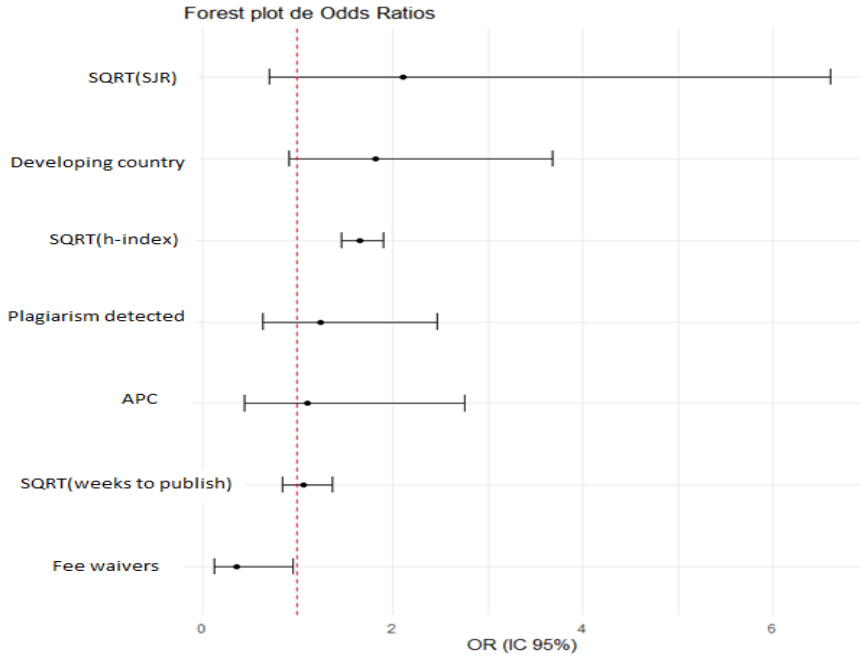

Figure 3: Forest Plot of the Model Coefficients With 95% Confidence Intervals.

Figure 4 shows comparative boxplots exploring the distribution of key quantitative variables (h-index, SJR, and publication week) by indexing status. Although some of these variables were not significant in the multivariate analysis, the plots revealed marked differences between active and excluded journals. Specifically, excluded journals tended to have lower h-index and SJR values, partially supporting the hypothesis that a lower bibliometric impact is associated with a higher risk of exclusion. These descriptive visualizations provided a helpful context for understanding the dataset structure and reinforcing the patterns inferred from the model. Together, both figures not only serve an illustrative function, but also offer complementary visual evidence that enhances the understanding, validity, and communicability of the results, in line with best practices in predictive modeling applied to editorial assessment.

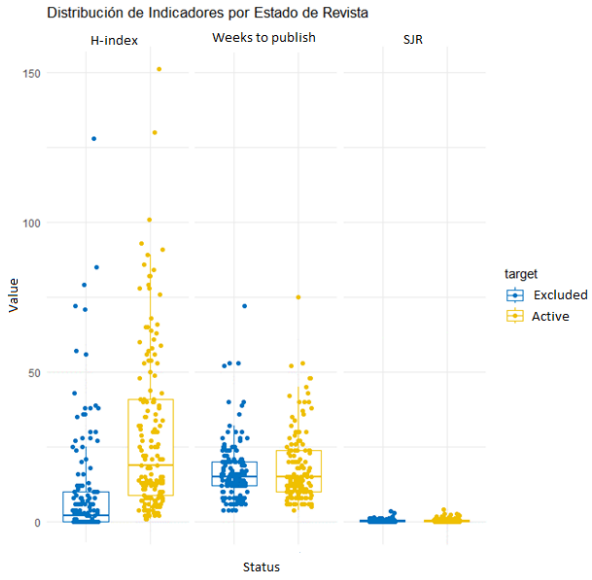

Figure 4: Comparative Box Plots of the H-Index, SJR and Weeks of Publication between Active and Excluded Journals.

3.5. Comparison with Alternative Models

As part of the complementary approach, two

additional models—LASSO regression and decision tree (CART)—were used to validate the consistency of the findings from the binary logistic regression. This comparison allows for the examination of both predictive performance and practical utility of different modeling approaches.

3.5.1. LASSO Regression

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a penalized technique that performs variable selection by applying a penalty to the absolute magnitude of coefficients (Zhou et al., 2024). This is particularly useful when certain variables have marginal or linear effects. In our model, implementation was conducted using the glmnet package in R with cross-validation to determine the optimal lambda value. The LASSO results confirmed that the h-index was the most robust variable, maintaining a nonzero coefficient even under conservative penalization levels. Variables such as "plagiarism detection," "APC charges," and "SJR" SJR were excluded from the model, aligning with their non-significance in the logistic regression.

3.5.2. Decision Tree (CART)

The classification tree was built using the CART (Classification and Regression Tree) algorithm via the rpart package. This technique offers explicit thresholds for decision-making (Sun et al., 2023). The tree identified the transformed h-index value as the first split, setting a threshold at $\sqrt{(\text{h-index})} < 0.5$ as indicative of higher exclusion risk. The second most relevant node was the absence of plagiarism detection policies, although with lower information gain.

Table 3: Model Comparison Metrics.

Model	AUC	Accuracy	Sensitivity	Specificity	Kappa
Logistic Regression	0.812	0.717	0.767	0.667	0.43
LASSO	0.803	0.708	0.755	0.661	0.41
Decision Tree	0.772	0.689	0.712	0.645	0.38

All three models converge on the h-index as the principal predictor, but logistic regression offers the best balance between sensitivity and specificity. While decision trees provide immediate interpretability for non-expert users, their overall performance is slightly inferior (Table 3). These comparisons reinforce the robustness of the logistic approach while validating the feasibility of alternative methods in contexts where transparency and rapid decision-making are priorities.

4. DISCUSSION

In recent years, the criteria for evaluation and retention in academic databases have evolved significantly due to the rise of open science, the demand for greater editorial transparency, and the strengthening of ethical standards in scientific publishing (Heen & Vogt, 2024). Open science has transformed traditional publishing by removing economic barriers and promoting the rapid dissemination and evaluation of scientific work (Penev, 2017). The Semmelweis University symposium analyzed the dangers of predatory journals, emphasizing the importance of thorough peer review and ethical practices to preserve scientific integrity (Benyó et al., 2024). Initiatives such as Plan S, promoted by cOAlition S, require publicly funded research to be published in journals that meet strict standards for open access, cost transparency, and responsible editorial practices (cOAlition S, 2018). Similarly, the TOP Guidelines have established criteria for evaluating journals based on their commitment to data openness, registration, and open peer review (Center for Open Science 2020). Regional platforms such as SciELO have adopted open science requirements since 2019, including author and reviewer identification, data sharing, and editorial transparency (Tonelli and Zambaldi, 2020; Lopes et al., 2023). Simultaneously, international databases such as Scopus and Web of Science have tightened their standards, excluding journals with deficiencies in peer review, editorial quality, or inconsistent metrics (Macháček & Srholec, 2022). Our study aligns with these trends by providing empirical evidence on the factors that influence the exclusion of journals from Scopus. A key finding is that fee waiver policies for authors from developing countries are negatively associated with journal retention, which may reflect financial limitations in journals promoting equity, but lacking economic sustainability. Databases, such as Scopus, present a structural bias against research from non-Western countries and publications in languages other than English. This bias may have led to the exclusion of legitimate journals from these regions, worsening their underrepresentation. Studies have highlighted the structural tendencies in indexing systems that may disproportionately affect journals from non-Western countries and those

publishing in languages other than English. For instance, Macháček and Srholec (2022) find cross-country disparities in the prevalence of potentially predatory journals indexed in Scopus, with higher rates observed in countries with limited editorial infrastructure. While not necessarily the result of explicit bias, these patterns suggest systemic challenges for journals from the Global South in meeting the indexing criteria. Such challenges include reduced access to funding, professional training, and international visibility, all of which can influence retention outcomes. Furthermore, excluding non-Western journals from major databases can hinder global innovation and reduce epistemic diversity by limiting the visibility and impact of research from these areas (Tennant, 2020). On the other hand, the h-index has been confirmed as a robust predictor of journal retention, consolidating its status as an indicator of prestige, impact, and scientific maturity (Ruscio, 2016). These results suggest that exclusion decisions are not solely based on ethical or inclusive criteria but also on a journal's ability to maintain stable bibliometric metrics. The adoption of the COPE, DOAJ, and OASPA guidelines could strengthen editorial governance and reduce exclusion risks (Silva & Moussa, 2024). COPE provides a detailed code of conduct and guidelines to address ethical issues, such as duplicate publication and authorship misconduct, which are essential for preserving the integrity of scientific publishing (Wager, 2012). Thus, this study offers tools for designing evidence-based science policies that align with the principles of integrity and openness. Journal exclusion is not merely a technical process but part of a broader debate on editorial quality in a saturated ecosystem threatened by predatory practices such as rapid and low-cost publishing (Soundarapandian, 2023; Xia et al., 2017). The h-index, as a key predictor, allows the development of objective monitoring models to distinguish structurally challenged but legitimate journals from those with questionable practices. Its relevance remains even after applying transformations to improve the model linearity, which is consistent with previous studies (Mongeon and Paul-Hus, 2016). The negative association between fee-waivers and journal retention should be interpreted with caution. While not questioning their inclusive value, this highlights the financial vulnerability of journals lacking institutional support, as noted by Shen and Björk (2015). The marginally positive trend observed in journals from developing countries suggests progress in professionalization, although not statistically significant, and aligns with an increase in the number and quality of Latin American journals (Delgado-Troncoso and Fischman 2014). Variables such as the SCImago Journal Rank (SJR), anti-plagiarism policies, and APC charges were not significant, possibly due to standardization or low variability among the journals analyzed. SJR is influenced by various scientometric factors, such as the h-index and citable documents, which may not correlate directly with journal activity status (Jamali et al., 2014). Anti-plagiarism policies in prestigious journals tend to be uniform, leading to little variation in their application. This uniformity suggests that these policies have a limited impact as determinants of journal activity (Marina & Sterligov, 2021). Additionally, APC costs vary widely, but are not directly related to a journal's operational status. They are more associated with business models and publication accessibility than with activity status (Marina & Sterligov, 2021). The random undersampling approach balanced the dataset and improved the detection of excluded journals (AUC = 0.812), as recommended by He and Garcia (2009) and Buda et al. (2018). These findings highlight the importance of combining quantitative metrics with structural factors to assess editorial sustainability. Indexing agencies offer tools for preventive strategies and technical support, particularly for journals committed to equity.

4.1. Implications for Scientific/Editorial Policy

Editors, particularly those operating in developing regions, must take a proactive approach to preserve their journals' indexing status by implementing structured monitoring systems. These systems should regularly track critical indicators, such as the h-index, publication timeliness, the rigor of peer review, and compliance with ethical guidelines, including anti-plagiarism policies and conflict-of-interest disclosures. The consistent tracking of these metrics enables editorial teams to identify performance gaps and implement targeted improvements before reaching critical thresholds that may trigger exclusion from indexing databases, such as Scopus. For journals that offer article processing fee waivers to authors from low- and middle-income countries, it is essential to establish long-term financial strategies that ensure that editorial operations remain sustainable. While fee waivers meaningfully contribute to equitable knowledge dissemination, they must not jeopardize the journal's financial and operational stability. This may involve seeking institutional subsidies, national science funding, or international partnerships to

maintain high editorial standards, while fulfilling inclusive mandates. This study supports the development of early warning systems based on editorial and bibliometric indicators from the perspective of indexing agencies. Such systems would alert journals at risk of exclusion and offer constructive data-driven recommendations for improvement. Prioritizing technical support and editorial capacity-buildingrather than relying solely on punitive delistingwould better align with the principles of scientific equity, especially in underrepresented regions. Furthermore, regional networks such as SciELO, Latindex, and Redalyc could serve as strategic allies in the implementation of these systems, providing contextualized support for journals operating in resource-constrained environments. Policymakers and academic institutions should recognize the importance of editorial quality as a dimension of scientific infrastructure and allocate resources accordingly. Overall, strengthening the sustainability, monitoring, and support structures of academic journals is essential to fostering a more inclusive, transparent, and resilient global publishing ecosystem. To ensure that early warning systems are effective and equitable, their implementation must be adapted to the reality of resource-constrained environments. In this context, regional indexing bodies such as SciELO, Latindex, and Redalyc can play pivotal roles by acting as intermediary support platforms. These organizations have deep contextual knowledge and established networks that can facilitate the localization of editorial standards, training programs, and diagnostic tools. For example, SciELO's continuous evaluation framework includes performance metrics and compliance checklists, which can be extended as part of an early alert system. Moreover, collaborative efforts between international indexing agencies and regional consortia can provide technical assistance, automated dashboards, or benchmarking tools tailored to the capabilities of small editorial teams. Such systems would allow journals to monitor indicators such as h-index stability, peer-review turnaround time, and transparency compliance, triggering non-punitive supportive interventions before delisting occurs. This approach reinforces inclusion, while promoting sustainable editorial practices across diverse publishing ecosystems.

4.2. Study Limitations

The sample, based on Scopus, SJR, and DOAJ data, focuses on open access journals, which limits the generalizability of results to subscription-based or hybrid models. The cross-sectional design does not allow for analysis of temporal dynamics, and random undersampling reduced the sample size, affecting statistical power. Future studies could compare editorial models, adopt longitudinal designs, or use advanced methods such as random forest.

5. CONCLUSION

This study introduces a novel application of predictive modeling to evaluate the risk of journal exclusion from Scopus, addressing a key gap in the literature on scientific publishing governance. By operationalizing editorial, bibliometric, and contextual variables through logistic regressionand validating the findings with LASSO and decision tree models, this research offers a data-driven framework to anticipate and mitigate exclusion risks in indexing systems. Among these findings, the h-index emerged as the most robust predictor of journal retention, underscoring its role in signaling editorial stability and scholarly impact. In contrast, journals offering fee waivers for authors from developing countries were more likely to be excluded, revealing a tension between equity and economic sustainability in open-access publishing. While not statistically significant, a positive trend among journals from developing regions suggests incremental improvements in professionalization and indexing compliance. Variables such as SJR and plagiarism policies were non-significant, possibly because of standardization across journals. The methodological use of random undersampling enhanced the predictive capacity of the model (AUC = 0.812), supporting its potential as a practical monitoring tool. Editors can apply this framework to proactively assess journal performance, while indexing agencies may use it to design early warning systems and support mechanisms, especially in collaboration with regional networks such as SciELO, Latindex, and Redalyc. Future research should explore the longitudinal dynamics of exclusion, incorporate citation behaviors and disciplinary differences, and apply advanced models, such as random forest or XGBoost. Comparative analyses with other indexing systems (e.g., Web of Science, DOAJ) could further strengthen the generalizability and utility of this predictive approach for global science governance.

REFERENCES

Anderson, C., Lowman, E. B., Doyle, S., & Sutton, D. (2022). Colonial and post-colonial history: Enhancing knowledge, capacity and networks in the Caribbean, sub-Saharan Africa and South Asia. LIAS Working Paper Series, 9(1), Article 1. https://doi.org/10.29311/lwps.202294105

Benyó, Z., et al. (2024). Scientific integrity in the era of predatory journals: Insights from an editors in chief symposium. British Journal of Pharmacology, 181(15), 2387–2390. https://doi.org/10.1111/bph.16480

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Center for Open Science. (2020). TOP Guidelines. https://www.cos.io/initiatives/top-guidelines

Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Coalition S. (2018). Plan S: Making full and immediate open access a reality. https://www.coalition-s.org

Delgado-Troncoso, J. E., & Fischman, G. E. (2014). The future of Latin American academic journals. In B. Cope & A. Phillips (Eds.), The future of the academic journal (2nd ed., pp. 379–400). Chandos Publishing. https://doi.org/10.1533/9781780634647.379

Elsevier. (2023). Content policy and selection criteria for Scopus. https://www.elsevier.com/solutions/scopus/how-scopus-works/content

Grudniewicz, A., et al. (2019). Predatory journals: No definition, no defence. Nature, 576(7786), 210–212. https://doi.org/10.1038/d41586-019-03759-y

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heen, E., & Vogt, H. (2024). Scientific rot: Unsustainable publishing practices threatens trust in medicine. Journal of Evaluation in Clinical Practice, 30(6), 941–944. https://doi.org/10.1111/jep.13989

Jamali, J., et al. (2014). Factors affecting journal quality indicator in Scopus (SCImago Journal Rank) in obstetrics and gynecology journals: A longitudinal study (1999–2013). Acta Informatica Medica, 22(6), 385–388. https://doi.org/10.5455/aim.2014.22.385-388

Lopes, P., et al. (2023). Adoption of open science criteria in SciELO.pt. BiblioCanto, 9(2), Article 2. https://doi.org/10.21680/2447-7842.2023v9n2ID33797

Macháček, V., & Srholec, M. (2022). Predatory publishing in Scopus: Evidence on cross-country differences. Quantitative Science Studies, 3(3), 859–887. https://doi.org/10.1162/qss_a_00213

Marina, T., & Sterligov, I. (2021). Prevalence of potentially predatory publishing in Scopus on the country level. Scientometrics, 126(6), 5019–5077. https://doi.org/10.1007/s11192-021-03899-x

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics, 106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Pandita, R., & Singh, S. (2023). Compromising quality parameters lead to fallout: A study of de-indexing of research journals. Malaysian Journal of Library and Information Science, 28(2), Article 2. https://doi.org/10.22452/mjlis.vol28no2.2

Penev, L. (2017). From open access to open science from the viewpoint of a scholarly publisher. Research Ideas and Outcomes, 3, e12265. https://doi.org/10.3897/rio.3.e12265

Quiroga-Garza, A., et al. (2022). Research barriers in the Global South: Mexico. Journal of Global Health, 12, 03032. https://doi.org/10.7189/jogh.12.03032

Ruscio, J. (2016). Taking advantage of citation measures of scholarly impact: Hip hip h index! Perspectives on Psychological Science, 11(6), 905–908. https://doi.org/10.1177/1745691616664436

Severin, A., & Low, N. (2019). Readers beware! Predatory journals are infiltrating citation databases. International Journal of Public Health, 64(8), 1123–1124. https://doi.org/10.1007/s00038-019-01284-3

Shen, C., & Björk, B.-C. (2015). 'Predatory' open access: A longitudinal study of article volumes and market characteristics. BMC Medicine, 13(1), 230. https://doi.org/10.1186/s12916-015-0469-2

Silva, J. A. T. da, & Moussa, S. (2024). The COPE / DOAJ / OASPA / WAME principles of transparency and best practice in scholarly publishing: A critical analysis. Ethics in Progress, 15(1), Article 1. https://doi.org/10.14746/eip.2024.1.7

Soundarapandian, J. (2023). Predatory publishing practices. TNOA Journal of Ophthalmic Science and Research, 61(4), 379. https://doi.org/10.4103/tjosr.tjosr_83_23

Sun, J., Jiang, N., Sun, G., & Huang, W. (2023). Analysis of CART algorithms in data mining. In 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM) (pp. 548–553). IEEE. https://doi.org/10.1109/MLCCIM60412.2023.00088

Tennant, J. P. (2020). Web of Science and Scopus are not global databases of knowledge. European Science Editing, 46, e51987. https://doi.org/10.3897/ese.2020.e51987

Tonelli, M. J., & Zambaldi, F. (2020). Data ownership and open science. Revista de Administração de Empresas, 59, 372–373. https://doi.org/10.1590/S0034-759020190601

Valz Gris, A., Cristiano, A., & Pezzullo, A. M. (2024). Integrity and accountability in academic publishing: Trends and implications of paper retractions and journal delistings. European Journal of Public Health, 34(Suppl_3), ckae144.678. https://doi.org/10.1093/eurpub/ckae144.678

Wager, E. (2012). Why has COPE developed guidelines for cooperation between journals and research institutions? Medicinski Arhiv, 66, 220–221. https://doi.org/10.5455/msm.2012.24.140-141

Williams, J. W., et al. (2023). Shifts to open access with high article processing charges hinder research equity and careers. Journal of Biogeography, 50(9), 1485–1489. https://doi.org/10.1111/jbi.14697

Xia, J., Li, Y., & Situ, P. (2017). An overview of predatory journal publishing in Asia. Journal of East Asian Libraries. https://www.semanticscholar.org/paper/An-Overview-of-Predatory-Journal-Publishing-in-Asia-Xia-Li/97a1c1f27fec0e5b3281ed1f7658c2f49c218539

Zhou, D. J., Chahal, R., Gotlib, I. H., & Liu, S. (2024). Comparison of Lasso and stepwise regression in psychological data. Methodology, 20(2), Article 2. https://doi.org/10.5964/meth.11523