

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Decoding the Secrets of Environmental Journals: Unraveling Journal Rankings through Text Mining and Predictive Analysis

Helena Dinis Oliveira

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**DECODING THE SECRETS OF ENVIRONMENTAL JOURNALS:
UNRAVELING JOURNAL RANKINGS THROUGH TEXT MINING
AND PREDICTIVE ANALYSIS**

by

Helena Dinis Oliveira

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor / Co Supervisor: Prof. Bruno Damásio, PhD

July 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Helena Dinis Oliveira

Lisbon, July 14, 2023

DEDICATION

Ao meu Pai Lúcio e à minha Mãe Jacinta

Grata por todo o Amor,

Esse não é até ver.

ACKNOWLEDGEMENTS

Aos Professores Bruno Damásio, Sandro Mendonça e Carolina Vasconcelos, por toda a orientação, disponibilidade, estímulo e aprendizagem. Muito obrigada!

Aos amigos, pelo apoio, espírito positivo e descontração; pelas aventuras e pela confiança.

À minha família, em especial à Grande Equipa:

Aos meus pais pelo suporte inquebrável, por todo o carinho, cumplicidade e acompanhamento. Obrigada pelo exemplo de superação, trabalho, método e Amor.

Aos meus irmãos, Adriano e André, obrigada por sempre estarem onde e quando for preciso.

A esta Equipa, obrigada por vibrarem com cada conquista, mas também pelo porto seguro que sempre encontrei em vós.

Requer continuação, Nery

ABSTRACT

Scientific journals foster the dissemination and validation of research findings contributing to the acquisition of knowledge that drives evolution. This study examined 12 environmental journals, six high-ranked and six low-ranked journals, to unravel journal rankings currently calculated based on articles' number of citations. Therefore, we applied text mining techniques to read the articles and a supervised approach to predict the group of journals in which an article would be published based on the article's content-wise variables and authorship variables. The results indicate ease of predicting the target: the models correctly predict around 90% of the articles. Our analysis of the extracted rules from the chosen model reveals that the variables that better discriminate the two groups of journals are the number of authors, references, and figures. This study contributes to better defining the differences between high and low-ranked journals despite the number of articles' citations. Thus, we step forward to help researchers strategically ensure their submitted papers align with the patterns identified in the group of journals desired. This may increase articles' chances of acceptance while maximizing their research impact within the scientific community.

KEYWORDS

Machine Learning; Text Mining; Bibliometrics; Scientometrics; Scopus; Scientific Journals

INDEX

1. Introduction	1
2. Literature review	3
2.1. Context	3
2.2. Concerns.....	4
2.3. Innovation in bibliometrics and related studies	5
3. Data and methods	6
3.1. Criterion of Journals' Selection - SJR.....	6
3.2. Research Procedure	7
3.2.1. Data Acquisition	7
3.2.2. Data Pre-Processing.....	8
3.2.3. Feature Engineering	9
3.3. Library's performance on content-features acquisition	11
3.4. Library's structure	12
4. In-depth analysis	13
4.1. Countries' affiliations	13
4.2. Number of authors	13
4.3. Sponsors.....	15
4.4. Abstracts' readability.....	15
4.5. Titles' word frequency.....	16
4.6. Research dissemination.....	17
5. Predictive Model Approach.....	18
6. Results and discussion.....	20
6.1. Models' Performance	20
6.2. Interpretation of the results.....	21
7. Conclusions	24
8. Limitations and recommendations for future works.....	25
9. References.....	26
Appendix	32

LIST OF FIGURES

Figure 3.1 - Criterion of journals' selection.	6
Figure 3.2 - Selected journals by country of the publisher.	6
Figure 3.3 - Research Procedure Diagram.	7
Figure 3.4 - Performance of the Data Gathering Process.	9
Figure 4.1 - Point percentage difference of countries' affiliations prevalence between high and low-ranked journals.	13
Figure 4.2 - Number of authors that have published in both groups of journals between 2016 to 2021.	14
Figure 4.3 - Median number of authors per journal and year of publication.	14
Figure 4.4 - Proportion of sponsored articles per journal.	15
Figure 4.5 - Abstracts' median readability value per journal and year of publication.	16
Figure 4.6 - Word cloud of the title's words.	16
Figure 4.7 - Mean value of the number of tweets and retweets that mention articles per journal and year of publication.	17
Figure 6.1 - Mean ROC curve for each model with its respective AUC and standard deviation, using repeated stratified 5-fold.	21
Figure 6.2 - Decision Tree rules.	22
Figure A.1 - Median value of the number of captures per journal and year of publication.	33
Figure A.2 - Median value of the number of citations per journal and year of publication.	34
Figure A.3 - Box plots for outliers' detection.	34
Figure A.4 - Spearman correlation matrix with outliers' removal.	35
Figure A.5 - Feature Selection: Features' importance from ANOVA test (without outliers).	35
Figure A.6 - Feature Selection: Features' importance from a Decision Tree (with outliers).	36
Figure A.7 - Feature Selection: Features' importance from a Decision Tree (without outliers).	36

LIST OF TABLES

Table 3.1 - Created variables' description.	10
Table 3.2 - Proportional Stratified Random Sampling: results on the correctness of number of figures and tables identified by the tool.	11
Table 4.1 - Descriptive statistics on the number of publications by author in each group of journals.	14
Table 6.1 - Performance metrics for each model.	20
Table 6.2 - Features' relative importance on each model.	23
Table 6.3 - Percentage of mislabelled articles, per journal, in the test set: results from the Decision Tree model.....	23
Table A.1 - Journals' metadata.	32
Table A.2 - Types of documents considered as articles per journal.	33

LIST OF ABBREVIATIONS AND ACRONYMS

1_NE	Nature Energy
2_EES	Energy and Environmental Science
3_AEM	Advanced Energy Materials
4_ACSEL	ACS Energy Letters
5_NS	Nature Sustainability
6_ESM	Energy Storage Materials
7_EPSE	Environmental Progress and Sustainable Energy
8_ESARUEE	Energy Sources, Part A: Recovery, Utilization and Environmental Effects
9_IJAE	International Journal of Ambient Energy
10_TS	Thermal Science
11_APJCE	Asia-Pacific Journal of Chemical Engineering
12_IJSDP	International Journal of Sustainable Development and Planning
F1	F1-score
RESE	Renewable Energy, Sustainability and the Environment
ROC curve	Receiver Operating Characteristic curve
SJR	Scimago Journal Rank
WoS	Web of Science

1. INTRODUCTION

As the world seeks development and innovation, publishing research articles in journals enables the scientific breakthrough that drives evolution. Knowledge acquisition is part of a collaborative process so that different perspectives are considered to enhance the results. Thus, journals play an essential role in making available knowledge between scientists and promoting the dissemination of the discovered findings (Masic, 2021), emerging as the “*vehicles of communication between scholars and scientists*” (Herubel, 1999, p. 1). Therefore, scientific journals have a preponderant impact on follow-up scientific research.

Journals’ prestige is nowadays perceived through scientometric indicators generally based on the number of citations of the documents published. However, one may question if the number of citations is a reasonable approach to classifying journals’ quality. Are citation counts an indicator of research articles’ quality? We identify a self-perpetuation loop: reputed journals receive more attention, which results in more citations of documents on those journals, leading to a higher scientometric indicator, and then the loop starts again. Therefore, despite its quality, it is difficult for an article not published in a high-ranked journal to be cited and known.

Therefore, this study aims to answer the following research question: What hidden factors contribute to the stratification of environmental journals?

We strive to explore patterns that distinguish research articles between high and low-ranked journals while characterizing the authors and research environment in which they were developed, such as authors’ affiliations and funding. Thus, we take a data science-guided approach by applying text mining techniques, descriptive and in-depth analysis, and a predictive approach.

As research objectives, we highlight the followings:

1. To build a library that automatically and consistently acquires the data.
2. To extract content-wise variables from the articles’ text, such as the number of figures and tables.
3. To describe differences, if applicable, between high and low-ranked journals in the following dimensions: authorship-wise, content-wise, and in a scientometric perspective.
4. To identify the patterns that distinguish high and low-ranked journals from the articles’ content perspective as well as the research environment in which they were produced, if applicable.

This study will examine journals within the purview of *Renewable Energy, Sustainability and the Environment* (RESE), which significantly impacts today’s society due to the growing climate concerns (Valencia et al., 2019). As stated by Rosokhata et al. (2021), the publications in this area have increased, especially in the last decade and a half. Indeed, research in this scope has been seen as a hope to contribute to new solutions to energy demand, promoting the usage of viable and sustainable energy sources.

The contribution of this research is twofold. Firstly, we perform an in-depth analysis to overview today’s spectrum of research. Thereby, the fundamental themes to be explored are the research environment in which an article was developed and how the article is structured.

Secondly, we build supervised models to predict whether an article will be published in a high or low-ranked journal based solely on the themes mentioned.

The remainder of this work is organized as follows. Section 2 exposes bibliometric and scientometric context as well as related studies. Section 3 introduces the methodology taken. Section 4 presents an in-depth data analysis. Section 5 covers a predictive approach to the data. Section 6 presents and discusses the results of the previous chapter. Section 7 concludes the study, while Section 8 indicates the study's limitations, proposing future approaches.

2. LITERATURE REVIEW

2.1. CONTEXT

Bibliometrics provides a set of tools to assess scientific production. This is a concept first introduced by a Belgian librarian (Otlet, 2021) in 1934, being defined as *“the measurement of all aspects related to the publication and reading of books and documents”* (Rousseau, 2014, p. 1). However, it is not consensual, and others consider Pritchard as the pioneer of the term being defined as *“the application of mathematics and statistical methods to books and other media of communication”* (Pritchard, 1969, p. 2). More recently, Spinak (1996) stated that bibliometrics studies the organization of the scientific and technological sectors from bibliographic and patent sources to identify actors, their relationships and trends. According to the latter, this field of study has increased immensely, having taken distinct approaches.

One can perform a bibliometric study on authorship. In this case, the researcher explores the existence of intellectual communities by evaluating how authors collaborate amongst themselves in publications and their respective institutional affiliations and geographic dispersion (Aistleitner et al., 2023; López-Robles et al., 2019). Moreover, we may also analyze on a sociodemographic level by exploring the academic background. Another interesting approach is to evaluate the topics and their evolution (Alvarez-Betancourt & Garcia-Silvente, 2014). Therefore, it is possible to understand how researchers adapted their interests as new research problems arise. Lastly, from a more content-guided point of view, the methodologies and techniques performed to develop the documents may also be analyzed.

Being a sub-field of Bibliometrics, Scientometrics was defined as the *“quantitative study of science, communication in science, and science policy”* (Hess, 1997, p. 75), measuring and analyzing scientific literature. These two areas ought to contribute to the assessment of scientific papers published in journals which are yearly evaluated by, amongst other indexes, the SCImago Journal & Country Rank (SJR) indicator, being a measure of journals' prestige.

In 2007, the SCImago research laboratory developed the SJR indicator, which uses data from Scopus - a database built in 1996. This is a scientometric indicator being calculated as *“the average number of weighted citations received in the selected year by the documents published in the selected journal in the three previous years, --i.e. weighted citations received in year X to documents published in the journal in years X-1, X-2 and X-3”* (SJR - Help, n.d.). To cover that not all citations are worth the same, the weights of the citations are attributed considering the importance of the journals the citations have come from. For the purpose mentioned, SCImago uses a similar algorithm as Google's PageRank (Ascaso, 2011). Therefore, when other prestigious journals cite an article, the weight of this citation increases in the computation of the index. The calculation of the impact factor index – IF - (another scientometric measure) does not apply this logic. IF also considers self-citations, contrary to the calculation method of SJR (Falagas et al., 2008).

In the scope of this study, the most prominent advantage of SJR over the Journal Citation Reports (JCR) impact factor index is that the former's portal (SCImago) is freely available online. In contrast, the latter's portal (JCR) is accessible if paid (Brown, 2011). Moreover, Scopus (SCImago database) has evidence to cover more journals than the Web of Science (Mongeon & Paul-Hus, 2016).

2.2. CONCERNS

As research progresses, scientometric indicators become more used to evaluate the quality of papers quantitatively, taking massive importance on the research paradigm. In other words, roughly, the higher the number of citations, the better the quality of the paper is perceived to be.

The ranking of journals researchers have published in is also a topic to consider when foundations provide research grants. As described by Khokhlov (2020, p. 2): *“to obtain high-quality scientific results, you need money, and you can get it only after publishing a significant number of ‘cool’ articles”* in reputed journals.

Furthermore, publications have a massive impact on the career progression of scholars (Chapman et al., 2019; Rawat & Meena, 2014), being the number of publications a decisive matter when universities recruit researchers, exposing the maxim *Publish or perish*. Ioannidis et al. (2018, p. 3) discussed this topic, *“hyperprolific authors (...) might also reflect idiosyncratic field norms, to say the least”*, referring to authors who have published, on average, 72 papers per year, between 2000 and 2016, in other words, an average of one paper every five days.

Indeed, some researchers and writers are skeptical about the reliability of an index based solely on citation count as a measure to rank journals. Taylor et al. (2008, p. 12) expose concerns about the issue mentioned above: *“the evaluation of science and scientists based on the JIF and other single-parameter indices is problematic, and further investigation needs to be done in order to design a better index to measure scientific quality”*.

Reinhard Werner, a highly respected physicist, claims an interesting point of view: when scholars cite one another, in fact, *“an ongoing pool is being conducted to elect the best quality papers”* (Werner, 2015, p. 1). A citation should refer to another work to invoke inspiration, identify previous work that was decisive for the present study, or criticize. Therefore, having more citations might not be a consequence of the excellent work the research has performed; on the contrary, it can reflect massive flaws. Moreover, Werner (2015, p. 1) claims that ethical issues may arise. Scholars may cite to: *“boost a friend’s citation statistics; to satisfy a potential big-shot referee; and to give the impression that there is a community interested in the topic by stuffing the introduction with irrelevant citations to everybody, often recycled from earlier papers”*.

Overall, the general public may perceive low-indexed journals as presenting less-quality ideas, data, language, mathematical processing of results, or the design of illustrations (Khokhlov, 2020). Even though the content may not follow the perceived reputation due to high citation counts, manuscripts that appear in high-ranked journals are automatically considered high-quality documents.

A concrete experience that may demonstrate the abovementioned problem is stated by Khokhlov (2020, p. 2): *“I wanted to know how many animals the authors used in their experiments and how they analysed their survival curves. Naturally, this determines the significance of results. [...] I did not find such information either in the article itself or in the Supplementary Materials published on the journal’s website. And this is an article that has been cited over 1800 times”*. Thus, the assessment of the article's quality may not have prioritized the transparent disclosure of the approach taken.

Therefore, it is difficult to perceive the quality of the journals based on citations, given all the ethical constraints mentioned above. As a result, we aim to partially analyze the articles' content and take insights that contribute to a different perspective on environmental research journals.

2.3. INNOVATION IN BIBLIOMETRICS AND RELATED STUDIES

Owing to the ongoing fast evolution of the Internet, the dissemination of knowledge has changed profoundly (Baheti & Bhargava, 2017). Therefore, Altmetrics takes a step further with a more complete and updated approach to evaluating research through the creation of alternative metrics to the traditional citation count. Altmetrics aims at perceiving the articles' social impact and visibility by measuring how it circulates on the Internet through multiple online sources, e.g., social networks, blogs or repositories (García-Villar, 2021), unraveling the impact of research in other less explored dimensions. After Elsevier acquired Plum Analytics in 2017, a more general view of how people interact with research documents was acquired, creating PlumX metrics (use case of Altmetrics), which will be examined in this project.

Antons et al. (2020) analyzed 124 text-mining research articles, using Web of Science to retrieve articles' metadata. Researchers manually read the articles and extracted important information, such as the software and algorithms used for the analysis and the transparency and replicability of the studies. We aim to extract content-related variables in an automated manner by analyzing a more significant number of articles than the study mentioned above.

Thelwall (2022) and Thelwall et al. (2023) concluded that using machine learning models, citations (as a proxy of articles' quality) can be predicted based on namely articles' metadata. In this study, we will not only consider metadata but also extract information from the articles' text, such as the number of figures and tables.

Similarly to the scope of our study, Brembs et al. (2013) compares high and low-ranked journals from the research health field and concludes that articles from high-ranked journals are less or no more *methodologically sound* than those published in low-ranked journals. The concept of methodological soundness is referred to as a measure replication having in mind the experimental and statistical rigor.

3. DATA AND METHODS

This section presents the study's methodology outlining the systematic approach used to gather data. Additionally, we describe the transformations performed on the data in order to derive more meaningful insights into the analyses conducted in Sections 4 and 5.

3.1. CRITERION OF JOURNALS' SELECTION - SJR

To select the journals under study, we took the approach described in Figure 3.1. Thus, we consider the journals based on their SJR supported by the discussion in Section 2.

Firstly, the list of the SCImago-indexed RESE journals from 2021 was extracted¹, consisting of 197 entries, excluding the book series, conferences and proceedings, and trade journals. Due to the usage of the Scopus database to access articles' metadata, it was imperative to check journal availability in this source of data, being the list reduced to 194 journals. In fact, only two journals were not found in Scopus, which highlights the high journal coverage of this database. To have a sufficient number of articles per journal to conduct the analyses, we restricted the list of journals to only the ones with no less than 150 documents² published in 2021. Thus, considering the top and bottom 10% of the journals from the resulting list of 65 journals, we dealt with the six best-ranked and the six worst-ranked journals. Table A.1 in the Appendix lists these journals along with detailed information that characterizes them.

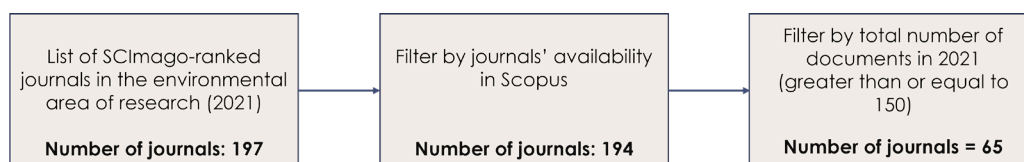


Figure 3.1 - Criterion of journals' selection.

The United Kingdom and the United States are the publishers' countries that dominate the spread of knowledge in the RESE field - Figure 3.2.

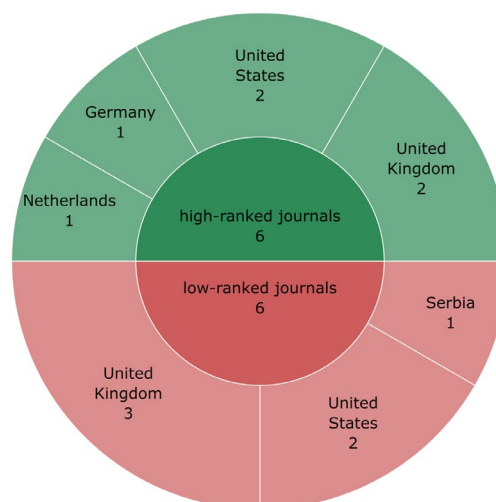


Figure 3.2 - Selected journals by country of the publisher.

¹ The data was acquired in September 2022.

² This threshold was defined from our judgment and perception.

3.2. RESEARCH PROCEDURE

The research procedure taken is presented in Figure 3.3 and explained in detail below.

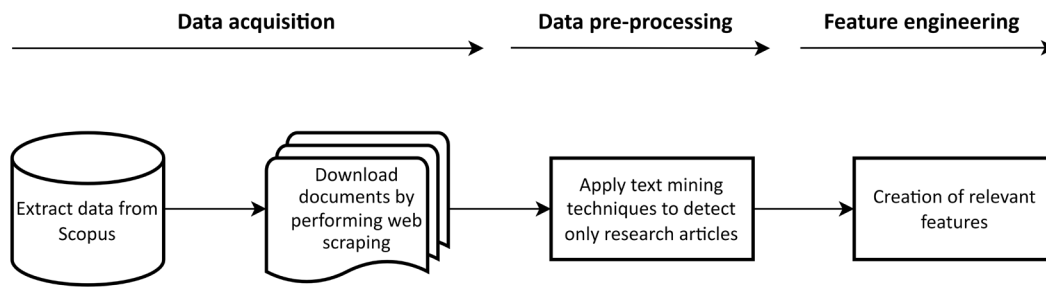


Figure 3.3 - Research Procedure Diagram.

3.2.1. Data Acquisition

3.2.1.1. Extract data from Scopus

Metadata about scientific documents must be available to measure scientific production from different perspectives adequately. Therefore, the data acquisition stage began with collecting articles' metadata³, a pillar for the study's success.

Scopus and Web of Science (WoS), owned by Elsevier and Clarivate, respectively, are the two most well-known databases (Mongeon & Paul-Hus, 2016), being both repositories of publications' metadata.

The latter was tested in a preliminary stage of the study and has demonstrated to have less code-friendly solutions to retrieve data from the database and proved to be very time-consuming in the acquisition of data.

The former was tested, having available a friendlier library both in data gathering and speed, *pybliometrics* (Rose & Kitchin, 2019). Apart from this, WoS does not differentiate authors with the same name (Pranckutė, 2021). In contrast, Scopus disambiguates authors by providing an identifier that distinguishes researchers, which is essential when developing analysis at an authorship level, as in this study.

Although both present disadvantages (Falagas et al., 2008; Visser et al., 2021), we decided to use Scopus for the scope of this project due to the advantages presented both above and in the Literature Review section that steered the choice of using SJR (a ranking index that uses data from this database).

Pybliometrics is a Python package that accesses the RESTful APIs provided by Scopus, revealing it as a handy tool. This package provides content-wise data, e.g., Title, Abstract, References, and Keywords, providing information about the tenor of the document. On an authorship note, the authors' names, respective affiliations, and foundation sponsors are also available. Lastly, it is also known the number of references and citations. The usage of this library allows for consistent data acquisition, is less prone to errors, and ensures the reproducibility of the results.

³ The data was acquired in February 2023.

To test the library developed, further explained, and better describe how journals' articles have changed over the years and compare characteristics of the articles from the two groups of journals (high and low-ranked), the time frame considered was from 2016 to 2021.

3.2.1.2. Web scraping tool to download the articles

The physical document acquisition process was performed through web scraping, which is a term that refers to the extraction of content and data from a webpage in an automated manner. Therefore, it was essential to fully understand how the webpage could provide the intended information; such logic would vary depending on the target website. It was imperative to download the documents to extract more relevant features to enrich our analysis.

3.2.2. Data Pre-Processing

3.2.2.1. Evaluate only research articles

Different types of documents may also hide different patterns since their characteristics vary. Being aware of this matter, we evaluated only research articles to establish an accurate comparison.

Although, at first glance, this information could be easily accessed from Scopus' metadata, the attribute that discriminates the document type was misleading. After downloading the articles, we realized that even though the data was filtered by type-article, other types of documents were also being considered. This happens because journals have different definitions of document types. This study considers research articles as original scientific work describing new experimental results or theoretical research findings. We gathered the terms that referred to this definition of research article in each journal and summarised this information in Table A.2.

Thus, since the aim is to consider only research articles, a text mining challenge arose. This consisted of identifying the articles' structure from each journal to correctly capture which type of document we were evaluating. This was a demanding task since the structure of the articles varies from journal to journal and even within the same journal. Having the information on the type of article, we validated the definition of document types of each journal accordingly.

It is essential to mention that the procedure to filter articles amongst the documents from Scopus also excluded the documents that could not be read by the tool because they were corrupted. Therefore, even though some journals were not a target of this filtering (due to the absence of documents other than articles), some documents may have been removed due to this matter.

The performance of the data-gathering process is presented in Figure 3.4. The tool has considered 14 841 articles from 2016 to 2021 out of the 17 087 documents transferred.

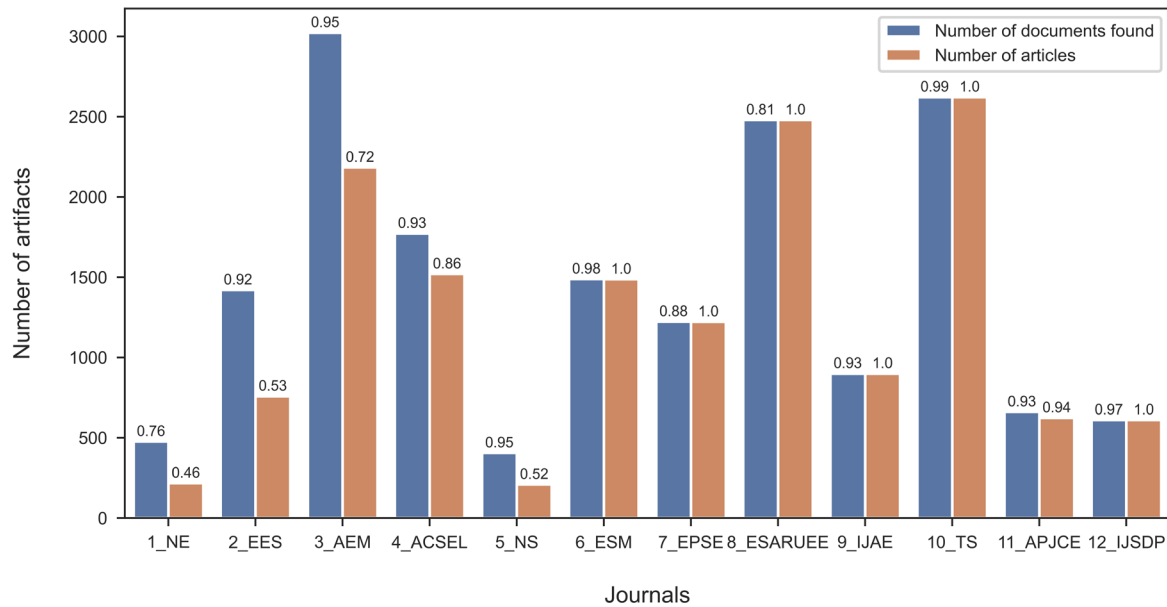


Figure 3.4 - Performance of the Data Gathering Process.

Note: The numbers above each blue bar represent the proportion of documents downloaded from the list of documents from Scopus. The numbers above each orange bar represent the proportion of articles among the documents.

The number of articles varies significantly amongst the different journals, ranging from about 250 to 2700. Regarding performance, Nature Energy was the journal with the lowest efficiency on document metadata acquisition (we only found 76% of the documents retrieved by Scopus). The Thermal Science journal had the highest success rate in the previous matter, being additionally the journal with the highest number of articles.

3.2.3. Feature Engineering

It is imperative to ensure that the features used to create new variables are in the expected format. Author affiliations identifiers were separated by ';' and '-', and the publication date was also not presented in a consistent format.

To achieve our third and fourth research objectives of exploring the articles from a content and research environment perspective, we were confronted with the need to formulate new variables that would enable a vast exploration of these themes. In fact, creating relevant features assumes a preponderant role in any data analysis or modeling to uncover interesting patterns. Measuring the ease of understanding of research articles has been proven useful in unraveling interesting patterns (Jamar, 2022; Yeung et al., 2018). Thus, we decided to explore this topic by using a readability index. Additionally, affiliations play an essential role in the production of scientific papers, therefore, we created variables under this scope that highlight, for example, affiliations' geographic dispersion. We also identified huge potential of content-wise features to reveal how authors disseminate their findings, such as the use of a visual or schematic representation of their findings or procedures (number of figures and tables).

Table 3.1 lists the variables created. Below, we present a more complete description of the created features.

Variable name	Description
flesch	Flesch-Kincaid readability index on the abstract.
gini_affiliations	Gini index on the diversity of affiliations.
gini_affiliations_country	Gini index on the diversity of affiliations' countries.
is_sponsored	Binary variable: 1 if the article has been sponsored; 0 otherwise.
nr_figures	Number of figures.
nr_tables	Number of tables.
nr_images	Number of images (number of figures + number of tables)
title_len	Number of characters in the title.
title_word_count	Number of words in the title.

Table 3.1 - Created variables' description.

3.2.3.1. Flesch-Kincaid readability index

Considering the articles' abstracts, we built a readability feature using the Flesch-Kincaid Grade Level index (Kincaid et al., 1975). It is an adaptation of Flesch Reading Ease (Flesch, 1948), one of the first metrics in this scope. The former retrieves a grading USA level of education required to comprehend a given text. They are both calculated considering the average number of words per sentence and the average number of syllables per word, assuming that longer sentences and words are more challenging to read. Shortly, the Flesch-Kincaid index is interpreted as the higher the score, the higher the level of USA education, therefore the more complex the understanding, being the formula presented below.

$$Flesch - Kincaid\ index = 0.39 * \frac{total\ words}{total\ sentences} + 11.8 * \frac{total\ syllables}{total\ words} - 15.59 \quad (1)$$

3.2.3.2. Gini - diversity of affiliations/affiliations' countries

A feature that accessed the diversity of affiliations and their corresponding countries was of interest. For this purpose, we computed a Gini index where a higher value indicates a greater diversity of authors' affiliations in a given article, providing more synergies. Analogously, we computed the index to the countries of each affiliation so that it is possible to evaluate the cooperation between authors from institutions of different countries.

The formula below calculates the Gini impurity. In the case of the Gini index's calculation on the affiliations of the articles, J is the set of affiliations of a specific article, and p_i the probability of choosing an affiliation with label i .

$$Gini\ impurity = 1 - \sum_{i=1}^J p_i^2 \quad (2)$$

3.2.3.3. Number of figures and tables

After format handling, we conducted a text-mining approach toward having as much information from the articles as possible. This consisted of automatically 'reading' the articles using regular expressions to extract the number of figures and tables presented. This process proved to be demanding due to the different articles' layouts and the exclusion of supplementary figures/tables.

3.3. LIBRARY'S PERFORMANCE ON CONTENT-FEATURES ACQUISITION

In order to assess the performance of the tool in extracting the number of figures and tables presented in the articles, we selected a random sample of articles using the formula presented below.

$$sample\ size = \frac{\frac{z^2 * p * (1 - p)}{e^2}}{1 + \frac{z^2 * p * (1 - p)}{e^2 * N}} \quad (3)$$

Where:

z – z-score (1.96)

p – standard deviation (0.5 – conservative approach)

e – margin of error

N – population size (14 827 articles)

Due to the unavailability of resources to check a sample of 996 articles resulting from a margin of error of 0.03, we analyzed 96 articles using a margin of error of 0.10.

Having the number of articles to analyze, we conducted a proportional stratified random sampling for each journal, meaning that the sample size of each stratum is proportional to its share in the population. Therefore, Table 3.2 shows the number of articles analyzed per journal (n_h), the total number of articles per journal (N_h) and the proportion of articles well assigned, by the tool, in terms of the number of tables and figures.

Journal abbreviation	n_h	N_h	$\hat{p}_h$ tables	$\hat{p}_h$ figures
1_NE	1	218	1	1
2_EES	5	759	1	1
3_AEM	14	2185	1	1
4_ACSEL	10	1521	1	0.90
5_NS	1	210	1	1
6_ESM	10	1488	1	0.90
7_EPSE	8	1223	0.63	0.50
8_ESARUEE	16	2480	1	1
9_IJAE	6	900	1	1
10_TS	17	2622	1	0.88
11_APJCE	4	624	1	1
12_IJSDP	4	611	1	1

Table 3.2 - Proportional Stratified Random Sampling: results on the correctness of number of figures and tables identified by the tool.

Using $\hat{p} = \sum \frac{N_h}{N} \hat{p}_h$, we conclude that the tool has correctly extracted the number of figures in 91.7% of the articles and the number of tables in 96.9%.

3.4. LIBRARY'S STRUCTURE

Python was the programming language chosen for the execution of the project owing to its versatility, practicality and ease of data manipulation to conduct a data science project. The data acquisition process, both from Scopus and from manipulating the pdf documents, required massive code. Thus, if not developing the code using a module's logic, the program's maintenance, reproducibility and efficiency are challenged.

Generally, writing code in an object-oriented logic relies on creating classes where properties, attributes and methods (in other words, functions) define the object *per se*, its behaviours and actions. One of the benefits of object-oriented programming is code organization and reusability (Lutz, 2013). These can be obtained due to the classes' inheritance feature. In other words, parent classes extend properties and behaviours to child classes, avoiding repeated code that may bring maintenance problems.

Therefore, the library developed is composed of a superclass that defines the general behaviour of any journal, and then each subclass, representing a specific journal, establishes how it is going to behave by overwriting methods defined in the superclass or by defining properties that characterize its journal.

It is important to mention that if the researcher desires to apply functionalities of this library to a non-explored journal, all it is needed to do is to create a new class along with the attributes and methods that specify the behaviour of the journal.

This library may help scientists to broaden this study to other areas of research or other journals under the renewable and environmental area by reproducing the methodology conducted in this project.

4. IN-DEPTH ANALYSIS

In this section, we conduct a descriptive analysis of the data from 2016 to 2021 guided through the point of view of two groups of journals: high and low-ranked journals.

4.1. COUNTRIES' AFFILIATIONS

Regarding authors' affiliation, Rosokhata et al. (2021) concluded that the topic of renewable energy is mainly affiliated by institutions from economically developed countries.

We took a different approach by studying the prevalence of countries' affiliations in each group of journals and then computed the difference in point percentage. We concluded that the most common affiliation country in each group is China and that more countries are playing a role in worst-ranked journals (127) than in best-ranked journals (116).

The world map from Figure 4.1 depicts the percentage point difference of the feature mentioned above between the high and low-ranked journals. Therefore, we can contrast the United States and India, being the former the affiliations' country that is more polarised in high-ranked journals and the latter more polarised in low-ranked journals.

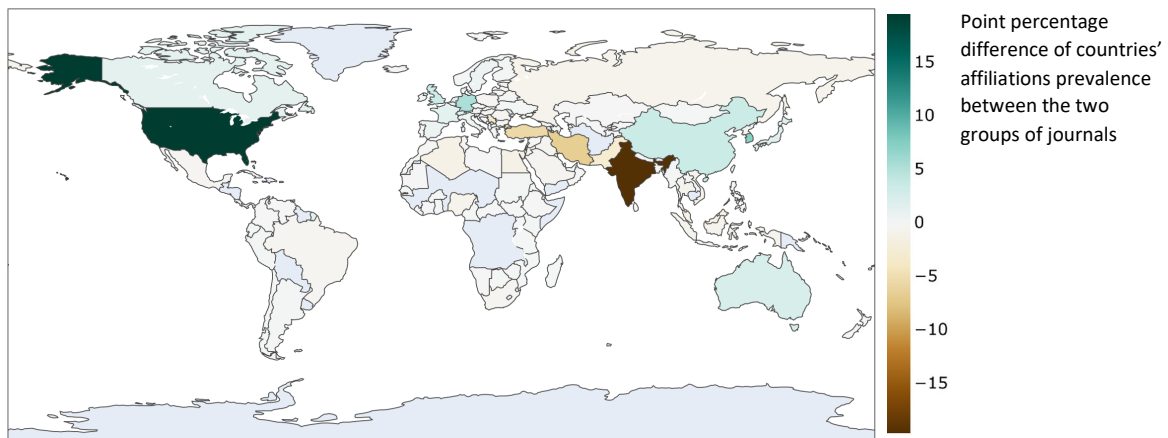


Figure 4.1 - Point percentage difference of countries' affiliations prevalence between high and low-ranked journals.

Note: In practical terms, it calculates each country's prevalence in each group of journals and takes the differences. A darker green represents a high positive difference (more present in high-ranked journals), and a darker brown a high negative difference in percentage points.

4.2. NUMBER OF AUTHORS

To leverage the author's disambiguation provided by Scopus, we aimed to understand the publication patterns of the authors across the two groups of journals. A total of 52 910 authors published RESE articles in the journals considered, 89 (about 0.17%) of them published both in high-ranked and low-ranked journals from 2016 to 2021.

An interesting approach that arose was to identify how many of the authors, that have published in both groups of journals, published in a low-ranked journal after a publication in a high-ranked journal (27 authors) - Figure 4.2. This sequence of events may demonstrate that researchers publish their work regardless of the journal's reputation. Similarly, note that 48

authors have published an article in a high-ranked journal after a publication in a low-ranked journal. Additionally, 14 authors took both these sequences of events.

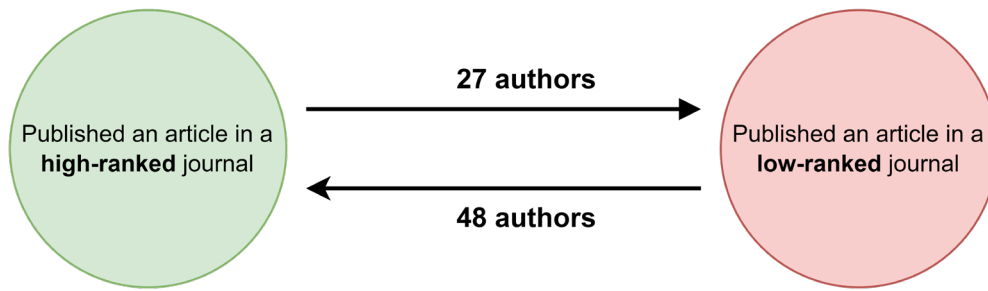


Figure 4.2 - Number of authors that have published in both groups of journals between 2016 to 2021.

To better understand in which group of journals authors publish more frequently we took a descriptive approach - Table 4.1. We conclude that authors that publish in high-ranked journals publish more times than those that publish in low-ranked journals, since the maximum, mean and third-quartile values are greater in this group.

Number of publications by author	Min	Max	Mean	Median	Percentile 75%	Std
High-ranked journals	1	64	1.77	1	2	2.09
Low-ranked journals	1	45	1.36	1	1	1.19

Table 4.1 - Descriptive statistics on the number of publications by author in each group of journals.

Additionally, in Brunson et al. (2017), where the topic explored was biomedicine, we conclude that more popular journals published higher-authorship articles, corroborated by the analysis of Figure 4.3.

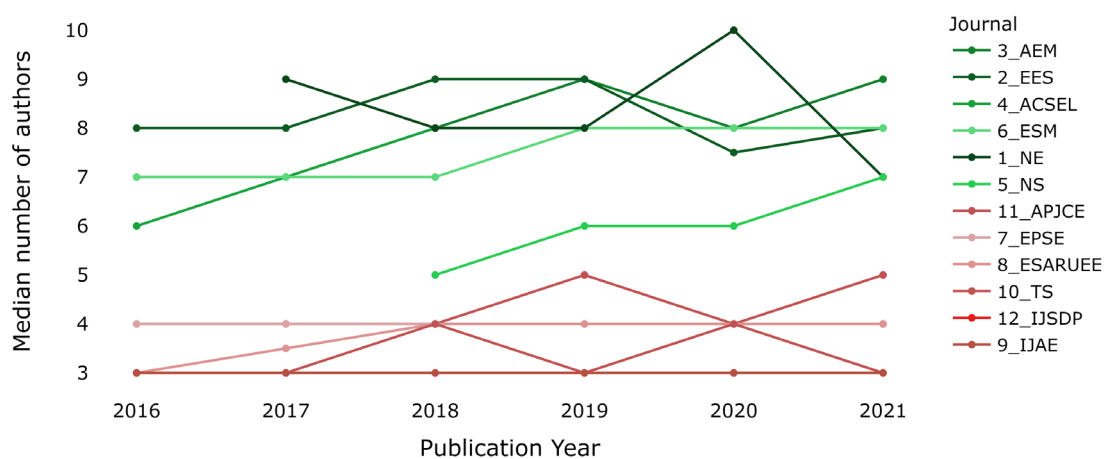


Figure 4.3 - Median number of authors per journal and year of publication.

Note: Some journals' median values overlap.

4.3. SPONSORS

Funding is crucial to obtain a high-quality research article since it usually provides the researcher access to specialized resources, such as advanced laboratory equipment, cutting-edge technologies, or research collaborations with experts in the field.

By exploring this theme, we identified a clear pattern: high-ranked journals have higher sponsored rates, over 0.8, while low-ranked journals have up to 60% of their articles sponsored- Figure 4.4. Thus, high-ranked journals' articles tend to be produced in an environment with more resources.

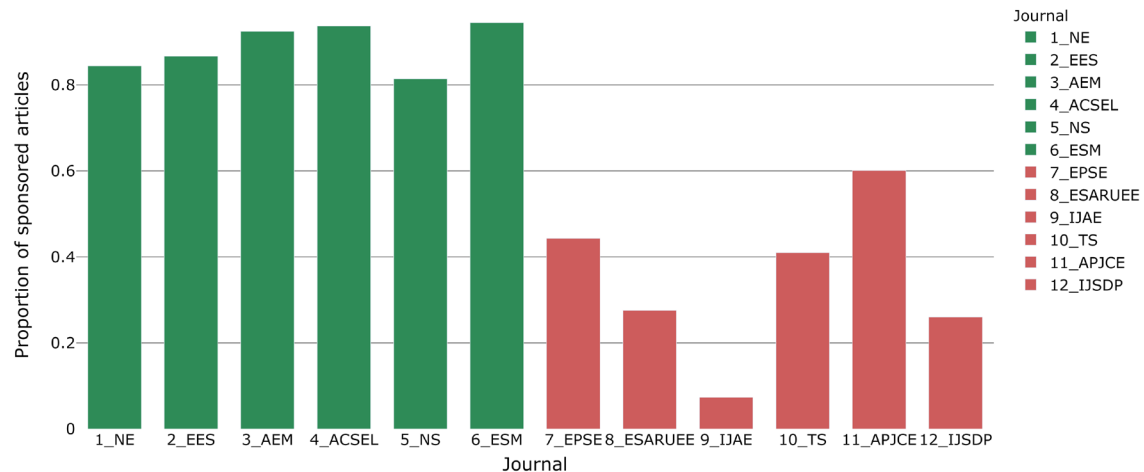


Figure 4.4 - Proportion of sponsored articles per journal.

4.4. ABSTRACTS' READABILITY

The proliferation of knowledge is also determined by how the message is delivered to the reader. Therefore, it is vital to present the study in a clear and easily understandable manner that facilitates the transfer of knowledge.

Thus, we explored the Flesch readability index of abstracts across time, aggregating the data by computing the median value for each year due to the presence of extreme values that would jeopardize the mean - Figure 4.5.

From a visual standpoint, it appears that the readability score in articles from less prestigious journals has more fluctuations in time. Generally, despite 12_IJSDP being well integrated into the group of prestigious journals, high-ranked journals have the highest readability scores (all above 14.5), meaning abstracts are more challenging to read than in low-ranked journals' articles. This finding is corroborated by previous studies (Jamar, 2022; Yeung et al., 2018).

Note: The most frequent words are more prominent. On the left are words from high-ranked journals' articles; on the right are words from low-ranked journals' articles.

4.6. RESEARCH DISSEMINATION

To comprehend how RESE research articles disseminate in the community we analyzed various metrics that enable a diverse overview. The capture metric, Figure A.1 in the Appendix, refers to the number of people who have added the article to their library, being greater in high-ranked journals than in low-ranked journals. The same conclusion applies to the number of citations, as expected, since the ranking group of a journal is determined, to a significant extent, by the number of citations its articles receive – Figure A.2.

Exploring the proliferation of research in social media, Figure 4.7, we conclude that Twitter is getting more importance in research since articles published in 2021 have more tweets than those published in 2019, for example. Additionally, high-ranked journals' articles have a higher footprint on Twitter. This corroborates Twitter's importance, analyzed by Chan et al. (2022).

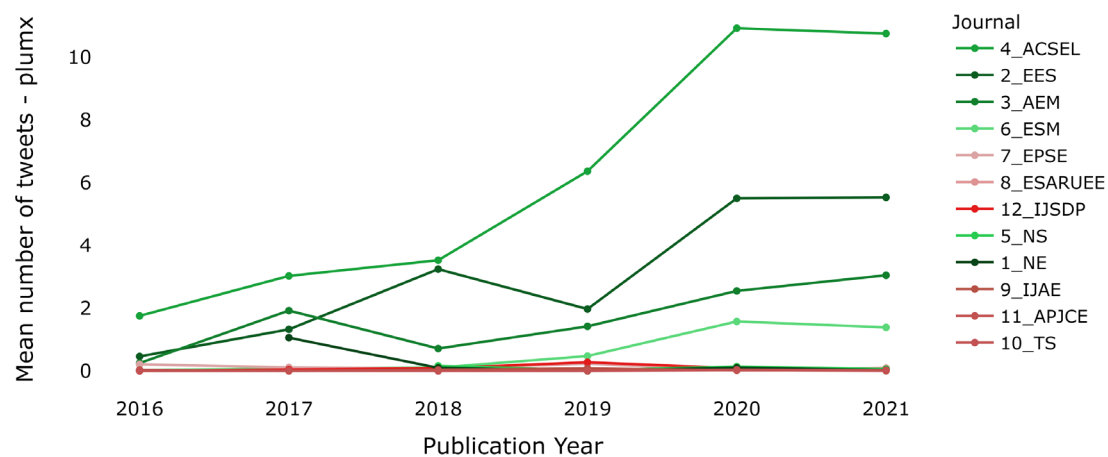


Figure 4.7 - Mean value of the number of tweets and retweets that mention articles per journal and year of publication.

5. PREDICTIVE MODEL APPROACH

In order to determine if content-wise variables alone can predict whether an article will be published in a high or low-ranked journal, we conducted a predictive approach to the data. In this study, we considered the SJR from 2021, calculated by considering the documents published in the three previous years. Owing to the way SJR is calculated, the articles analyzed were published from 2018 to 2020.

The quality of a model is highly affected by data quality (Ekström et al., 2021; He et al., 2019). Therefore, before any algorithm is performed, it is crucial to ensure the data is cleaned and in the correct format. Out of 8387 articles (45% of high-ranked journals; 55% of low-ranked journals), the only inconsistency encountered was an article without authors being excluded. The correct format of the variables was ensured before the analysis described in Section 4.

Subsequently, we proceeded with the data partition using the Hold-Out method. Thus, the data was split into a training set (60%) to fit the model, a validation set (20%) to tune the hyperparameters and control overfitting and a test set (20%) to access the quality of the model in unseen data. We split the data so that each created dataset maintains the same proportion of each journal's articles as the original dataset.

In order to detect the presence of extreme values, which have a major impact on some models' performance (Kalisch et al., 2016), we proceeded with the analysis of the variables' boxplots, removing 0.2% of the training data - Figure A.3. We took a simple approach to detect outliers so that we would have full control and interpretability on which articles were to be removed.

Feature selection significantly impacts the predictive model's performance, aiming at reducing redundancy among the features used and selecting variables with a discriminative power to the target feature (Punlumjeak & Rachburee, 2015). To support the chosen set of features, we checked, with and without outliers' removal various approaches. We ran Spearman correlations, Figure A.4, to guarantee the absence of correlated pairs of variables and, consequently, repeated information. This correlation also considers non-linear relationships between the variables, in contrast to the Pearson correlation (Hauke & Kossowski, 2011). Additionally, we ran ANOVA tests, Figure A.5, and a Decision Tree to check feature importance, Figure A.6 and Figure A.7, and tested different set combinations in each model.

In the Decision Tree and Random Forest models, we used the following features: *title_word_count*, *Flesch*, *gini_affiliations_country*, *nr_images*, *nr_references* and *author_count*. These are *non-normalized* features, using data without outliers' removal, due to the robust nature of the algorithms to extreme values.

In all the other models, due to their sensitivity to extreme values, the outliers were removed. Afterwards, we created logarithmic features in non-normalized features and applied a min-max scaler (in the train, validation and test sets) due to the sensitivity of scaling of the models used. The set of variables considered by these models were: *title_word_count*, *log_Flesch*, *log_nr_tables*, *log_gini_affiliations_country*, *log_nr_references*, *log_author_count*.

Regarding the supervised algorithms, it was a deliberate choice to employ a diverse set of predictive approaches. In fact, as no previous literature has lean over the matter studied in this project, we chose to take a wide spectrum of algorithms that extract information from the data in different ways. Thus, we performed a Decision Tree which recursively partitions the

data based on different attributes providing human-interpretable rules. Additionally, we ran an ensemble model, Random Forest, that takes into consideration the predictions of multiple decision trees. We also used a connectionist model, a Neural Network inspired by the structure and functioning of biological neural networks and a Logistic Regression based on the application of a sigmoid function, an S-shaped curve. Moreover, K-Nearest Neighbours, as a distance-based algorithm, relies on the concept that similar observations (calculated by a distance metric) are more likely to have similar outcomes. Lastly, we performed the Support Vector Machine algorithm which analyzes and classifies data by finding the optimal hyperplane in a high-dimensional feature space. Each employed algorithm possesses its own particularities, using different approaches to reveal patterns from the data. Therefore, it increases the chances of reaching a model with a favorable performance.

6. RESULTS AND DISCUSSION

6.1. MODELS' PERFORMANCE

To tune the hyperparameters for each model, we applied a grid search to exhaustively try different sets of parameters. This approach enabled the discovery of the set of parameters that maximizes the performance of the models.

The formulas of the metrics used to assess each model's performance are represented below. The abbreviation TP stands for True Positive, TN for True Negative, FP for False Positive, FN for False Negative, and T for total number of records.

$$Accuracy = \frac{TN+TP}{T} \quad (4)$$

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

In Table 6.1, we present performance metrics both in train and validation (to assess overfitting).

Model	F1 Train	F1 Validation	Accuracy Train	Accuracy Validation
Random Forest	0.912	0.910	0.920	0.917
Neural Network	0.900	0.895	0.909	0.903
Support Vector Machine (SVM)	0.910	0.895	0.917	0.902
K-Nearest Neighbours	0.903	0.888	0.912	0.897
Logistic Regression	0.897	0.886	0.905	0.895
Decision Tree	0.860	0.853	0.871	0.864

Table 6.1 - Performance metrics for each model.

The small difference in performance, of each model, between the train and validation datasets is an indicator of the models' capacity to generalize the patterns discovered, in contrast to fitting exactly against its training data - overfitting (Charilaou & Battat, 2022).

To seek a stable model evaluation, we applied a Repeated Stratified K-fold with five repeats. Thereby, the data was split into five folds: at each time, one-fold is used as a test set, and all the others are used to train the model; this process is repeated five times. Using different sets of training data and test data aims to reduce bias and variance (Awwalu & Ogwueleka, 2019; Raschka, 2018).

Figure 6.1 compares the performance between all the trained models and their power of discrimination, which is the degree of separability between the target classes. It presents each model's mean ROC curve (Receiver Operating Characteristic curve) out of the five folds.

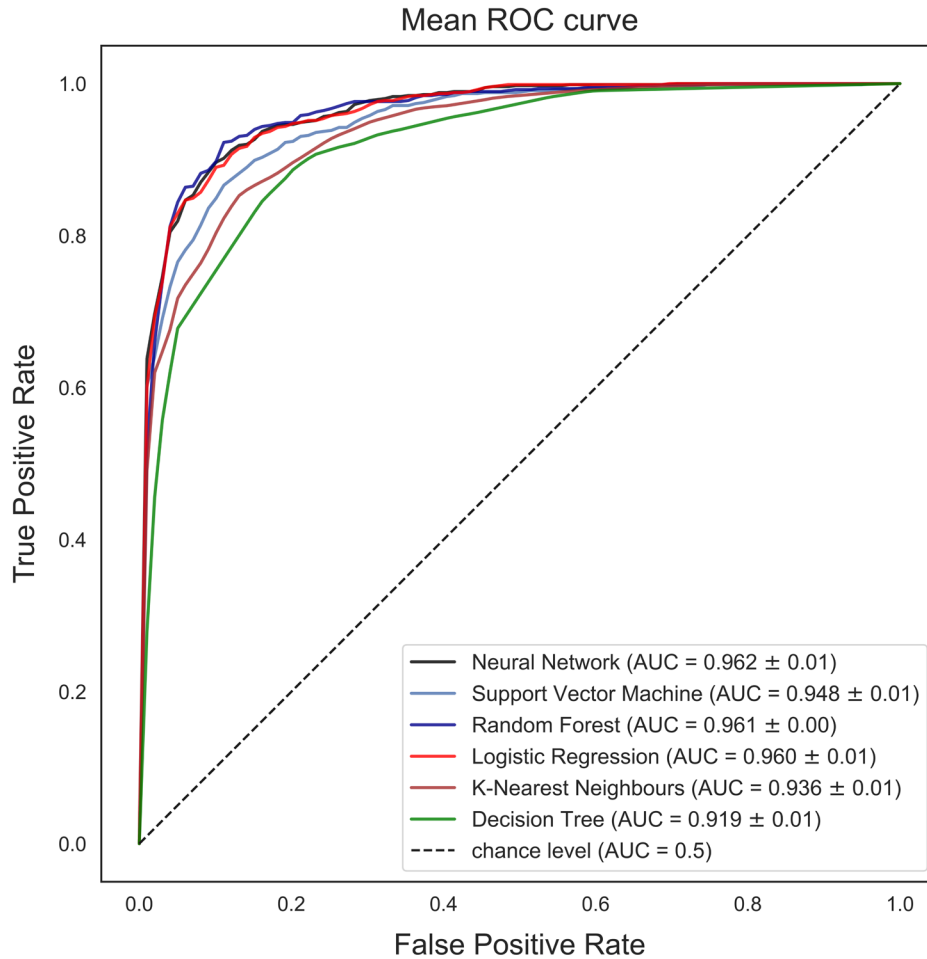


Figure 6.1 - Mean ROC curve for each model with its respective AUC and standard deviation, using repeated stratified 5-fold.

Contrarily to the results presented from the Hold-out method, the best-performed model from the Repeated stratified 5-fold was the Neural Network, followed by the Support Vector Machine.

6.2. INTERPRETATION OF THE RESULTS

Although the best-performed model is a Neural Network, the researcher must evaluate the trade-off between performance and interpretability, depending on the nature of the data and the purpose of the investigation. For example, in healthcare research, high performance is crucial, but knowing the rationale behind each prediction is also imperative to guide doctors in the final decision (Stiglic et al., 2020; Vellido, 2020). Interpretability is one of the advantages of simpler models when compared with more complex models, black-box models (Rudin & Radin, 2019), being complexity measured by the number of parameters computed by the model.

In this study, we opted to invest in the interpretability of the solution due to the urge to understand the patterns between an article's content-wise variables and the ranking group to which its journal belongs. From our perspective, the gain in performance of more complex

models does not offset their lack of interpretability. Thus, we will explore the results of the Decision Tree model.

One of the most critical parameters of a Decision Tree model is its maximum depth. In fact, as the tree's depth increases, the training dataset's performance increases as well. However, in the validation dataset, the performance may not follow this trend with the same magnitude, causing overfitting. Considering this, we pruned the tree early to mitigate the chances of overfitting.

Figure 6.2 reflects the rules the model has built to predict the ranking group of each article's journal. Therefore, we can infer the following simplified patterns from this rule-based algorithm:

- If an article has fewer than seven authors and fewer than 42 references, then it is predicted to be published in a low-ranked journal;
- If an article has more than six authors and fewer than 28 references, then it is predicted to be published in a low-ranked journal;
- If an article has more than six authors, more than 27 references and fewer than 13 images, then it is predicted to be published in a high-ranked journal.

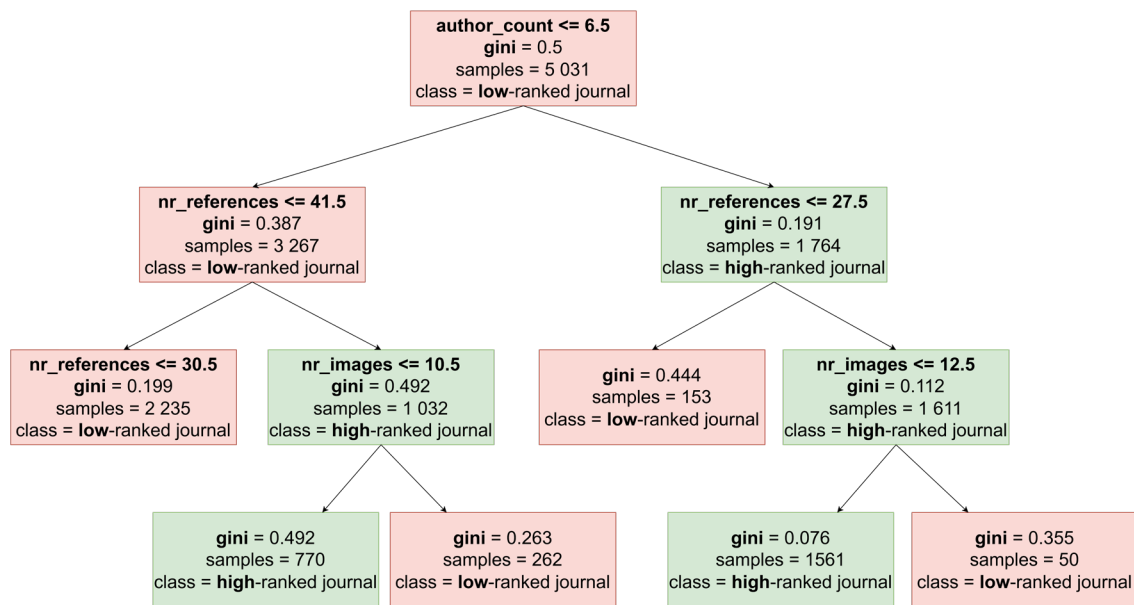


Figure 6.2 - Decision Tree rules.

Thus, the model only considered three variables: the number of authors, references and images. To understand the importance of every feature, we ran a Decision Tree without any depth restriction. In Table 6.2, we get the features' relative importance from the Random Forest (one of the best-performed models), from the pruned Decision Tree and, lastly, from the Decision Tree that has not been pruned. Following the three mentioned variables, the readability feature follows the rank, being the Gini feature on affiliations' countries the less important feature.

Features	Random Forest	Decision Tree not pruned	Decision Tree pruned
<i>author_count</i>	0.53	0.36	0.62
<i>nr_references</i>	0.25	0.27	0.28
<i>nr_images</i>	0.19	0.19	0.10
<i>Flesch</i>	0.03	0.11	
<i>title_word_count</i>	0.01	0.05	
<i>gini_affiliations_country</i>	~0	0.02	

Table 6.2 - Features' relative importance on each model.

We predicted the target for the test set (1 678 articles) defined in the beginning to understand better how the model would perform in unseen data. Thus, it was neither used for training nor in the validation set, ensuring independence in the assessment. The performance did not deviate from the one in the training and validation set, retrieving an F1-score of 0.864 and an accuracy of 0.863.

The highest-ranked journal out of the group of low-ranked journals is the one with more misclassified articles, in relative terms - Table 6.3. Therefore, it is the journal most similar to the group of high-ranked journals. Additionally, the journals Nature Energy and Nature Sustainability publish articles with similar characteristics due to the same proportion of mislabelled articles. From the test set, the model has correctly predicted 1 448 records (762 True Negative and 686 True Positive) and mislabelled 230 articles (153 False Positive and 77 False Negative).

Prediction	Journal abbreviation	Number of mislabelled articles	Total number of articles	Proportion of mislabelled articles
1	7_EPSE	77	140	0.550
1	12_IJSDP	11	51	0.216
1	8_ESARUEE	32	289	0.111
1	11_APJCE	6	69	0.087
1	10_TS	24	282	0.085
1	9_IJAE	3	84	0.036
0	4_ACSEL	29	165	0.176
0	6_ESM	28	200	0.140
0	1_NE	2	32	0.063
0	5_NS	2	32	0.063
0	3_AEM	13	261	0.050
0	2_EES	3	73	0.041

Table 6.3 - Percentage of mislabelled articles, per journal, in the test set: results from the Decision Tree model.

7. CONCLUSIONS

The tool developed contributed to the ease of data acquisition, meeting the first and second research objectives. To correspond to the third research objective, we highlight the major findings from the in-depth analysis.

By exploring author-related variables we conclude that a negligible percentage of authors (0.17%) have published in both high and low-ranked journals from 2016 to 2021. This indicates a clear stratification between authors who publish in high-ranked journals and those who publish in low-ranked journals. This can be explained by researchers' desire for greater visibility, prestige, and credibility associated with high-ranked journals. Moreover, high-reputed journals have articles produced by more authors, suggesting a collaborative nature of research, thus enhancing the results.

From a content-wise perspective, abstracts from articles published in high-ranked journals are more challenging to read, suggesting a higher level of complexity or technicality in their writing style. Additionally, articles' titles from high-ranked journals seem to be composed of more technical words, potentially reflecting the specialized research topics within these journals.

Regarding the dissemination of research, high-ranked journals' articles have a higher footprint on Twitter than articles published in low-ranked journals. This contributes to the rapid spread of the findings encountered by researchers and, consequently, may contribute to the higher number of citations owned by high-ranked journals.

In order to achieve the fourth research objective, we employed a supervised approach that yielded excellent results: the models correctly predicted the ranking group of the journals in around 90% of the articles. From the Decision Tree model, we concluded that the number of authors, references and images were the variables most important, and thus the ones associated to the stratification of articles between these two groups of journals.

Thus, it is possible to predict in which group of ranked journals an article will be published based on the content of the article and the research environment it has been produced (authors and affiliations-related variables). Thus, knowing the most relevant factors of environmental journals' stratification holds significant implications for the scientific community. In fact, through this study, we conclude that collaboration plays an essential role in publishing an article in a top environmental journal, highlighting the importance of knowledge exchange and the pooling of expertise. Moreover, the number of references reflects the research's depth, supported by previous literature, thus contributing to a more robust study. The number of images in an article contributes to a visual representation and communication of research findings, promoting more direct access to the information.

Thus, understanding the rules extracted from the Decision Tree model may help researchers to strategically manage their expectations, ensuring that their submitted articles align with the patterns identified for the articles from the ranking group desired. This may increase their chances of acceptance while maximizing their research impact within the scientific community.

8. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This study was conducted under the RESE area. Other areas of study may be explored in future research as well as the analysis of distinct types of documents other than research articles.

Interesting content-wise variables were left unexplored, such as authors' gender and the prevalence of references from the same journal the article has been published in. The former was difficult to acquire due to authors' names without a gendered tendency, mainly English or Asian names. The latter could not be explored due to a lack of data.

We also highlight the difficulty of accessing the actual quality of the research and its disruptiveness which should be a factor that contributes to the ranking of the journal in which an article has been published.

It would be interesting to conduct an unsupervised approach to explore the existence of clusters using articles' related features. Within each cluster, the articles would have similar characteristics, so a subsequent analysis of the ranking of their journals may unravel unexpected conclusions.

9. REFERENCES

- Aistleitner, M., Kapeller, J., & Kronberger, D. (2023). The authors of economics journals revisited: Evidence from a large-scale replication of Hodgson and Rothman (1999). *Journal of Institutional Economics*, 19(1), 86–101.
<https://doi.org/10.1017/S174413742200025X>
- Alvarez-Betancourt, Y., & Garcia-Silvente, M. (2014). An overview of iris recognition: A bibliometric analysis of the period 2000–2012. *Scientometrics*, 101(3), 2003–2033.
<https://doi.org/10.1007/s11192-014-1336-1>
- Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management*, 50(3), 329–351. <https://doi.org/10.1111/radm.12408>
- Ascaso, F. J. (2011). Factor de impacto, eigenfactor y article influence. *Archivos De La Sociedad Espanola De Oftalmologia*, 86(1), 1–2. <https://doi.org/10.1016/j.oftal.2010.12.005>
- Awwalu, J., & Ogwueleka, F. (2019). On Holdout and Cross Validation: A Comparison between Neural Network and Support Vector Machine. 6, 2394–9333.
- Baheti, A. D., & Bhargava, P. (2017). Altmetrics: A Measure of Social Attention toward Scientific Research. *Current Problems in Diagnostic Radiology*, 46(6), 391–392.
<https://doi.org/10.1067/j.cpradiol.2017.06.005>
- Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, 7.
<https://doi.org/10.3389/fnhum.2013.00291>
- Brown, T. (2011). Journal Quality Metrics: Options to Consider Other Than Impact Factors. *The American Journal of Occupational Therapy*, 65(3), 346–350.
<https://doi.org/10.5014/ajot.2011.001396>

- Brunson, J. C., Wang, X., & Laubenbacher, R. C. (2017). Effects of research complexity and competition on the incidence and growth of coauthorship in biomedicine. *PLOS ONE*, 12(3), e0173444. <https://doi.org/10.1371/journal.pone.0173444>
- Chan, C., Zeng, J., & Schäfer, M. S. (2022). Whose research benefits more from Twitter? On Twitter-worthiness of communication research and its role in reinforcing disparities of the field. *PLOS ONE*, 17(12), e0278840. <https://doi.org/10.1371/journal.pone.0278840>
- Chapman, C. A., Bicca-Marques, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., Guo, S., Hemingway, C. A., Leendertz, F., Li, B., Matsuda, I., Hou, R., Serio-Silva, J. C., & Chr. Stenseth, N. (2019). Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society B: Biological Sciences*, 286(1916), 20192047. <https://doi.org/10.1098/rspb.2019.2047>
- Charilaou, P., & Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, 28(5), 605–607. <https://doi.org/10.3748/wjg.v28.i5.605>
- Ekström, T., Burke, S., Wiktorsson, M., Hassanie, S., Harderup, L.-E., & Arfvidsson, J. (2021). Evaluating the impact of data quality on the accuracy of the predicted energy performance for a fixed building design using probabilistic energy performance simulations and uncertainty analysis. *Energy and Buildings*, 249, 111205. <https://doi.org/10.1016/j.enbuild.2021.111205>
- Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, 22(8), 2623–2628. <https://doi.org/10.1096/fj.08-107938>
- Falagas, M., Pitsouni, E., Malietzis, G., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal* :

- Official Publication of the Federation of American Societies for Experimental Biology*, 22, 338–342. <https://doi.org/10.1096/fj.07-9492LSF>
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233. <https://doi.org/10.1037/h0057532>
- García-Villar, C. (2021). A critical review on altmetrics: Can we measure the social impact factor? *Insights into Imaging*, 12(1), 92. <https://doi.org/10.1186/s13244-021-01033-2>
- Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30(2), 87–93. <https://doi.org/10.2478/v10117-011-0021-1>
- He, T., Yu, S., Wang, Z., Li, J., & Chen, Z. (2019). *From Data Quality to Model Quality: An Exploratory Study on Deep Learning* (arXiv:1906.11882). arXiv. <https://doi.org/10.48550/arXiv.1906.11882>
- Herubel, J.-P. (1999). Historical Bibliometrics: Its Purpose and Significance to the History of Disciplines. *Libraries & Culture*, 34.
- Hess, D. J. (1997). *Science Studies: An Advanced Introduction*. NYU Press.
- Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2018). Thousands of scientists publish a paper every five days. *Nature*, 561(7722), 167–169. <https://doi.org/10.1038/d41586-018-06185-8>
- Jamar, N. (2022). The readability of abstracts in library and information science journals. *Journal of Documentation*, ahead-of-print(ahead-of-print). <https://doi.org/10.1108/JD-05-2022-0098>
- Kalisch, M., Michalak, M., Sikora, M., Wróbel, Ł., & Przyszałka, P. (2016). Influence of Outliers Introduction on Predictive Models Quality. In S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, & D. Kostrzewa (Eds.), *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (pp. 79–93). Springer International Publishing. https://doi.org/10.1007/978-3-319-34099-9_5

- Khokhlov, A. N. (2020). How Scientometrics Became the Most Important Science for Researchers of All Specialties. *Moscow University Biological Sciences Bulletin*, 75(4), 159–163. <https://doi.org/10.3103/S0096392520040057>
- Kincaid, J. P., Fishburne, Jr., Robert P., R., Richard L., C., & Brad S. (1975, February 1). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*: <https://doi.org/10.21236/ADA006655>
- López-Robles, J.-R., Guallar, J., Otegi-Olaso, J.-R., & Gamboa-Rosales, N.-K. (2019). El profesional de la información (EPI): Bibliometric and thematic analysis (2006-2017). *Profesional de La Información*, 28(4), Article 4. <https://doi.org/10.3145/epi.2019.jul.17>
- Lutz, M. (2013). *Learning Python: Powerful Object-Oriented Programming*. O'Reilly Media, Inc.
- Masic, I. (2021). Bibliometric Indices and its Role for the Quality Assessment of the Authors Published Content in the Scientific Journals. *International Journal on Biomedicine and Healthcare*, 9(3), 244. <https://doi.org/10.5455/ijbh.2021.9.244-253>
- 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>
- Otlet, P. (2021). *Traité de documentation: Le livre sur le livre. Théorie et pratique*. In *Traité de documentation: Le livre sur le livre. Théorie et pratique*. Éditions des maisons des sciences de l'homme associées. <http://books.openedition.org/emsha/482>
- Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications*, 9(1), Article 1. <https://doi.org/10.3390/publications9010012>
- Pritchard, A. (1969). Statistical Bibliography or Bibliometrics? *Journal of Documentation*, 25, 348–349.

- Punlumjeak, W., & Rachburee, N. (2015). A comparative study of feature selection techniques for classify student performance. *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 425–429.
<https://doi.org/10.1109/ICITEED.2015.7408984>
- Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *ArXiv*. <https://www.semanticscholar.org/paper/Model-Evaluation%2C-Model-Selection%2C-and-Algorithm-in-Raschka/eef183687fab4d762a381f2e80e357e08e923f0a>
- Rawat, S., & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*.
<https://www.semanticscholar.org/paper/Publish-or-perish%3A-Where-are-we-heading-Rawat-Meena/f919c263974a914f662892445135159f9f586dd4>
- Rose, M. E., & Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, 10, 100263.
<https://doi.org/10.1016/j.softx.2019.100263>
- Rosokhata, A., Minchenko, M., Khomenko, L., & Chygryn, O. (2021). Renewable energy: A bibliometric analysis. *E3S Web of Conferences*, 250, 03002.
<https://doi.org/10.1051/e3sconf/202125003002>
- Rousseau, R. (2014). Forgotten founder of bibliometrics. *Nature*, 510(7504), Article 7504.
<https://doi.org/10.1038/510218e>
- Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1(2).
<https://doi.org/10.1162/99608f92.5a8a3a3d>
- SJR - Help*. (n.d.). Retrieved 8 January 2023, from <https://www.scimagojr.com/help.php>
- Spinak, E. (1996). *Diccionario Encicloédico de Bibliometría, Cienciometría e Informetría*.

- Taylor, M., Perakakis, P., & Trachana, V. (2008). The siege of science. *Ethics in Science and Environmental Politics*, 8(1), 17–40. <https://doi.org/10.3354/esep00086>
- Thelwall, M. (2022). Can the quality of published academic journal articles be assessed with machine learning? *Quantitative Science Studies*, 3. https://doi.org/10.1162/qss_a_00185
- Thelwall, M., Kousha, K., Wilson, P., Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., & Cancellieri, M. (2023). Predicting article quality scores with machine learning: The UK Research Excellence Framework. *Quantitative Science Studies*, 1–27. https://doi.org/10.1162/qss_a_00258
- Valencia, G., Núñez Alvarez, J., & Acevedo, C. (2019). Research Evolution on Renewable Energies Resources from 2007 to 2017: A Comparative Study on Solar, Geothermal, Wind and Biomass Energy. *International Journal of Energy Economics and Policy*, 9, 242–253. <https://doi.org/10.32479/ijeep.8051>
- Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112
- Werner, R. (2015). The focus on bibliometrics makes papers less useful. *Nature*, 517(7534), Article 7534. <https://doi.org/10.1038/517245a>
- Why the title of your paper matters. (2021). *Nature Human Behaviour*, 5(6), Article 6. <https://doi.org/10.1038/s41562-021-01152-2>
- Yeung, A. W. K., Goto, T. K., & Leung, W. K. (2018). Readability of the 100 Most-Cited Neuroimaging Papers Assessed by Common Readability Formulae. *Frontiers in Human Neuroscience*, 12. <https://doi.org/10.3389/fnhum.2018.00308>

APPENDIX

Journal	ISSN	SJR (2021)	SJR Rank	Publisher	Country	Coverage	Open Journal
Nature Energy	2058-7546	16.736	1	Springer Nature	United States	2016-2021	0
Energy and Environmental Science	1754-5706	11.558	2	Royal Society of Chemistry	United Kingdom	2008-2021	0
Advanced Energy Materials	1614-6840	8.226	3	Wiley-VCH Verlag	Germany	2011-2021	0
ACS Energy Letters	2380-8195	7.362	4	American Chemical Society	United States	2016-2021	0
Nature Sustainability	2398-9629	5.789	5	Nature Publishing Group	United Kingdom	2018-2021	1
Energy Storage Materials	2405-8297	4.865	6	Elsevier BV	Netherlands	2015-2021	0
Environmental Progress and Sustainable Energy	1944-7450	0.489	113	John Wiley & Sons Inc.	United States	2009-2021	0
Energy Sources, Part A: Recovery, Utilization and Environmental Effects	1556-7230	0.432	121	Taylor and Francis Ltd.	United Kingdom	2006-2021	0
International Journal of Ambient Energy	2162-8246	0.421	123	Taylor and Francis Ltd.	United Kingdom	1980-2021	0
Thermal Science	0354-9836	0.377	135	Vinca Inst Nuclear Sci	Serbia	2007-2021	1
Asia-Pacific Journal of Chemical Engineering	1932-2135	0.343	139	Wiley-Blackwell	United States	2006-2021	0
International Journal of Sustainable Development and Planning	1743-7601	0.284	148	International Information and Engineering Technology Association	United Kingdom	2006-2021	1

Table A.1 - Journals' metadata.

Journal Abbreviation	Type of documents filtered
Nature Energy	Articles
Energy and Environmental Science	PAPER; ARTICLE; (without classification)
Advanced Energy Materials	FULL PAPER; RESEARCH ARTICLE; RESEARCH PAPER
ACS Energy Letters	Letter
Nature Sustainability	Articles
Energy Storage Materials	X
Environmental Progress and Sustainable Energy	X
Energy Sources, Part A: Recovery, Utilization and Environmental Effects	X
International Journal of Ambient Energy	X
Thermal Science	X
Asia-Pacific Journal of Chemical Engineering	Research article
International Journal of Sustainable Development and Planning	X

Table A.2 - Types of documents considered as articles per journal.

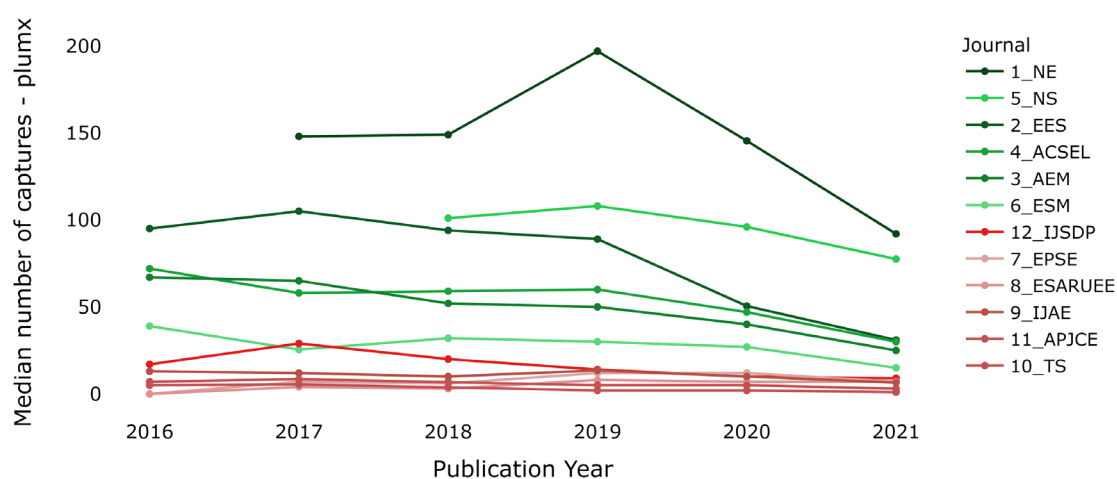


Figure A.1 - Median value of the number of captures per journal and year of publication.

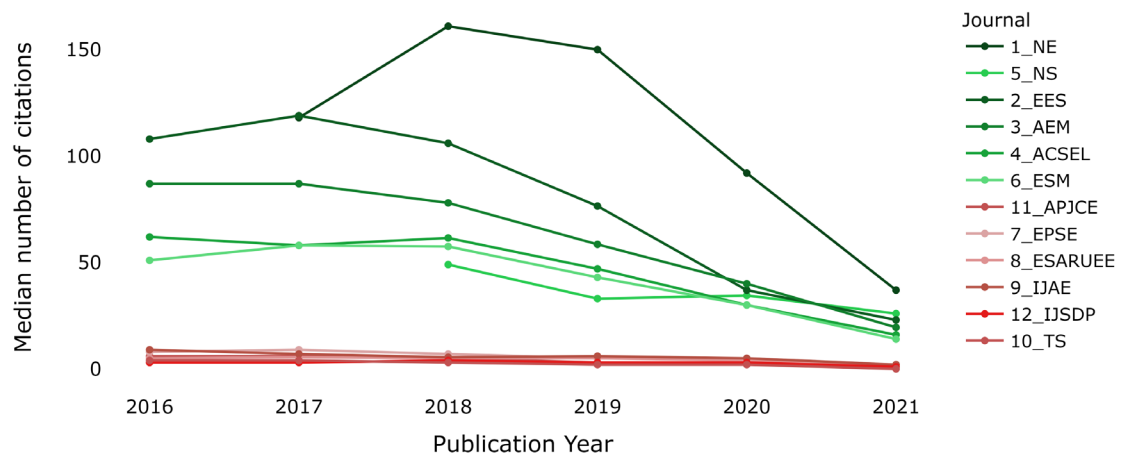


Figure A.2 - Median value of the number of citations per journal and year of publication.

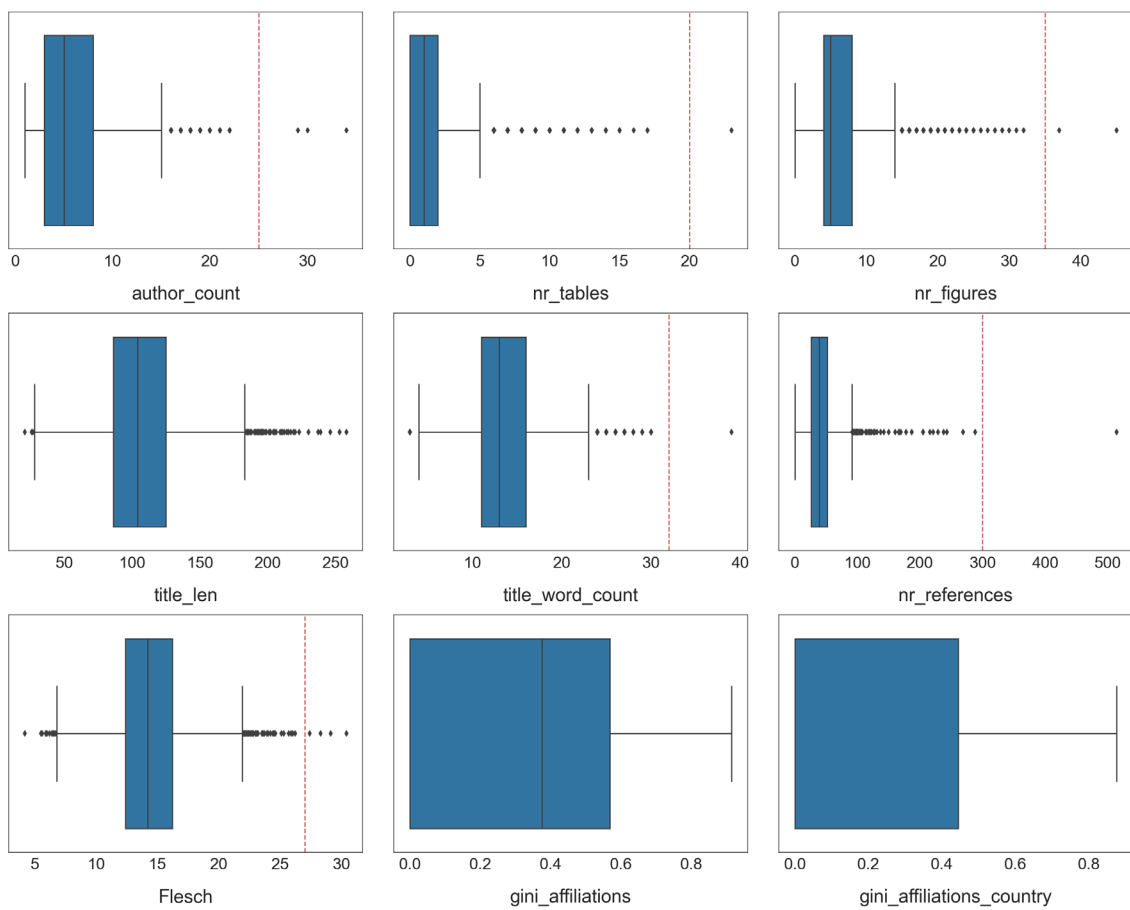


Figure A.3 - Box plots for outliers' detection.

Note: The red dashed line is the threshold used.

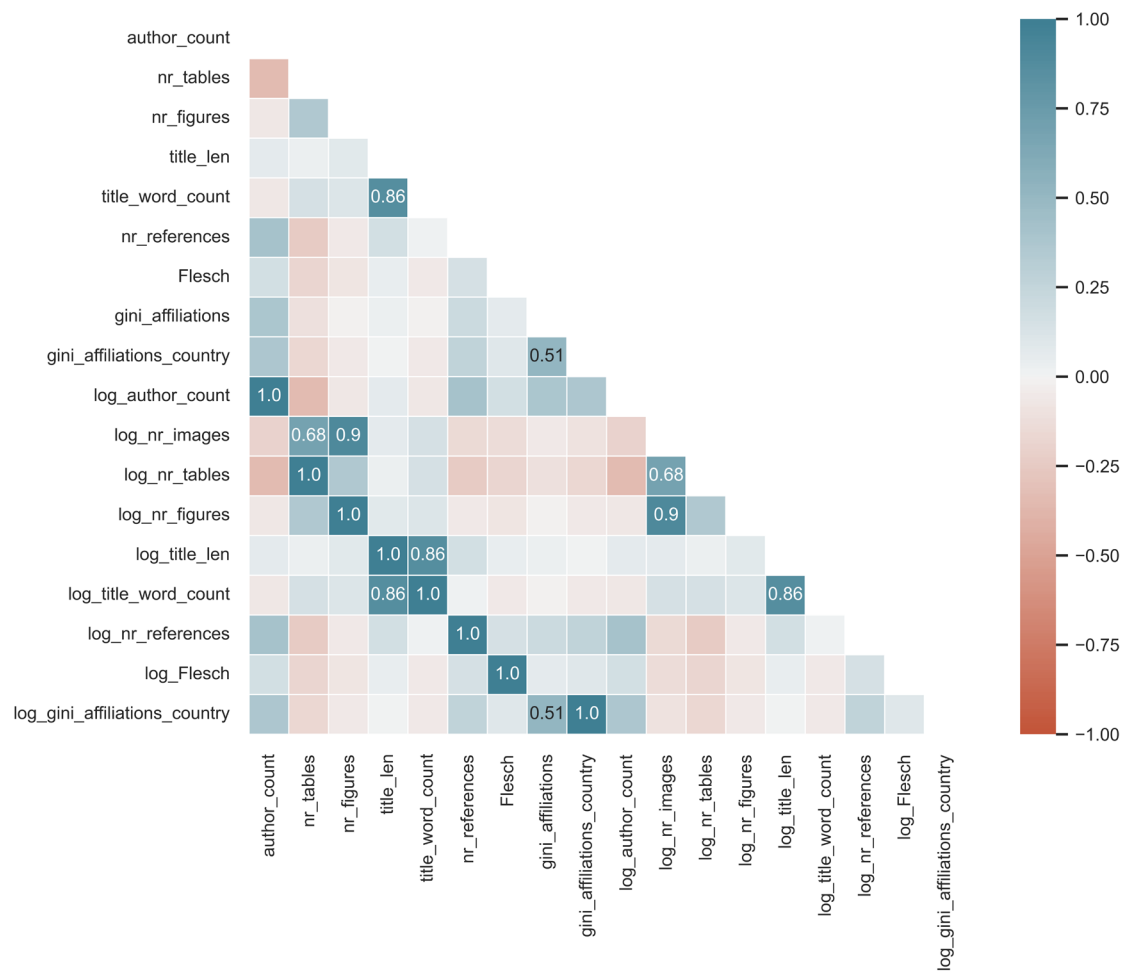


Figure A.4 - Spearman correlation matrix with outliers' removal.

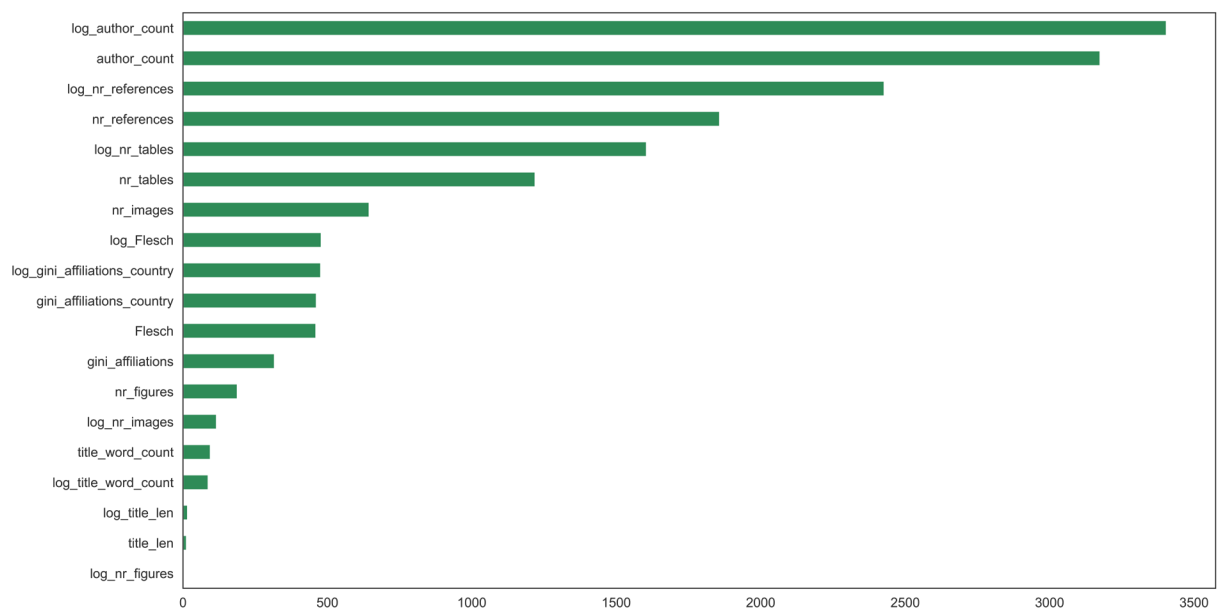


Figure A.5 - Feature Selection: Features' importance from ANOVA test (without outliers).

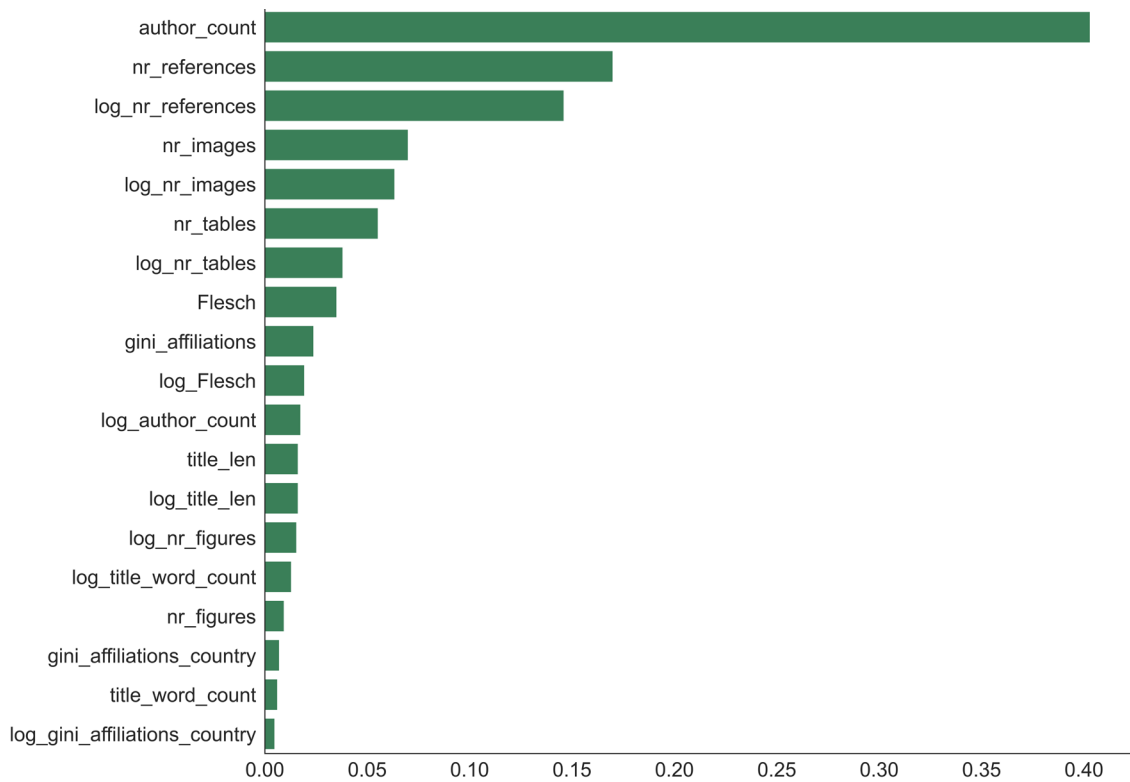


Figure A.6 - Feature Selection: Features' importance from a Decision Tree (with outliers).

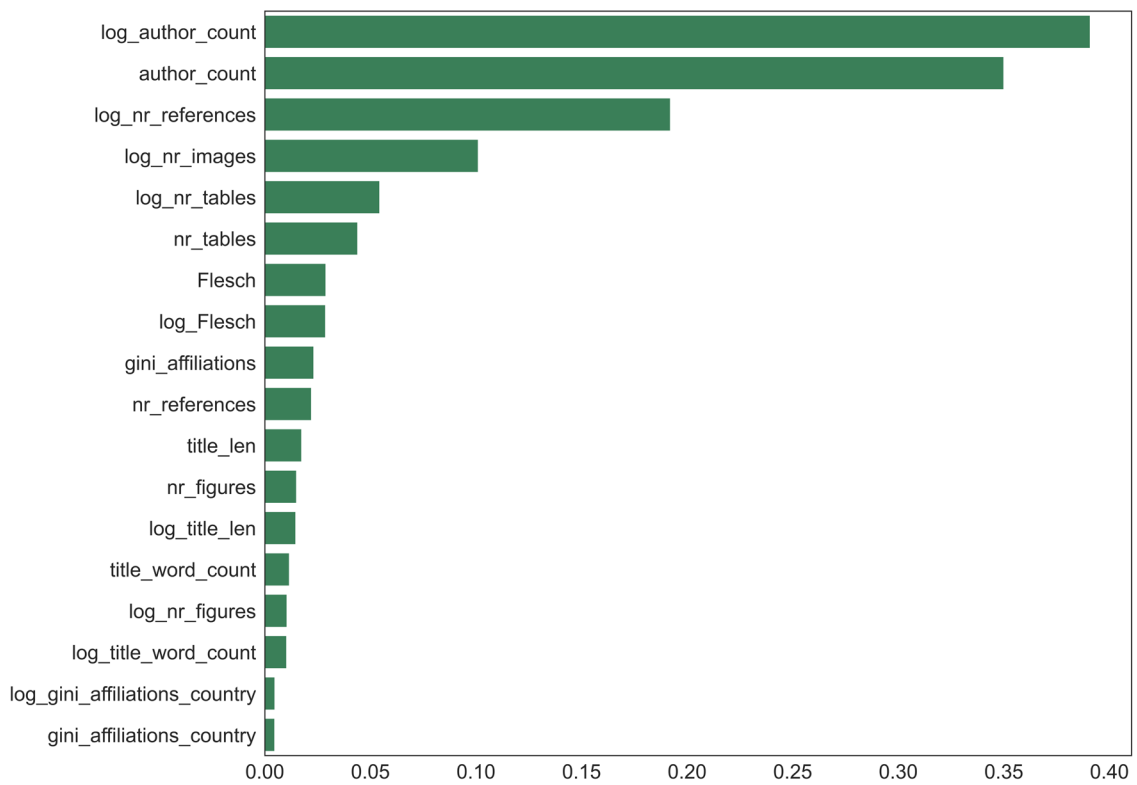


Figure A.7 - Feature Selection: Features' importance from a Decision Tree (without outliers).