IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 27, 2020, accepted November 10, 2020, date of publication November 16, 2020,
date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037665

A Contrast Pattern-Based Scientometric Study of the QS World University Ranking

OCTAVIO LOYOLA-GONZÁLEZ[1], MIGUEL ANGEL MEDINA-PÉREZ[2], RAYMUNDO ADRIÁN CORONILLA VALDEZ[2], AND KIM-KWANG RAYMOND CHOO[3], (Senior Member, IEEE)

[1]Tecnologico de Monterrey, Campus Puebla, Puebla 72453, Mexico
[2]Tecnologico de Monterrey, Campus Estado de Mexico, Ciudad López Mateos 52926, Mexico
[3]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Corresponding author: Miguel Angel Medina-Pérez (migue@tec.mx)

ABSTRACT Despite the shortcomings and criticisms of world university rankings, such metrics are widely used by students and parents to select institutions and by educational institutions to attract talented students and researchers, as well as funding. This article introduces the first contrast pattern-based scientometric study of world university rankings. Specifically, this study collects a database containing 34 features, which describe the essential research indicators for the top 200 universities in the Quacquarelli Symonds (QS) ranking. The use of 18 state-of-the-art classifiers in this database shows that the top 100 universities in the QS World University Rankings are separable from the remaining compared universities, achieving an average accuracy of 71%. Additionally, using a contrast pattern mining algorithm, a set of patterns describing the top 100 universities is extracted based on scientometric features. Additionally, this study proposes an approach for visualizing the extracted patterns to facilitate the decision-makers, such as senior university managers, in formulating and evaluating their research (ranking) strategies.

INDEX TERMS University rankings, data mining, scientometrics, contrast patterns.

I. INTRODUCTION

Scientometric studies have been shown to be an effective instrument for quantifying various metrics by analyzing the information extracted from the related data(sets) [1]. For example, using journals indexed by public databases (e.g., Scopus[1] and the Web of Science[2]), one can extract quantifiable criteria associated with the studied publications for analysis [2], [3].

The potential of scientometric studies is partly evidenced by the increasing number of related publications – see Fig. 1. For example, there are at least 300 scientometric publications as of 2019. The interest is not surprising due to their ability to display essential indicators, which can be used to evaluate and benchmark institutional research activities in the context of this paper. This, in turn, helps inform institutional and funding agency (agencies) decision-making.

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

[1]www.scopus.com
[2]https://clarivate.com/webofsciencegroup/solutions/web-of-science/

Scientometric studies can potentially help universities to identify their weaknesses and formulate strategies to improve their research indicators [1]. Additionally, these rankings are tools for prospective students and their parents in their university application(s). Universities can also use their rankings to increase their attractiveness to talented students, researchers, and research funding agencies [4], [5].

As shown in Fig. 1, there are several scientometric studies reported in the literature, including those relating to university rankings [2], [3], [6]–[8]. These university rankings generally take into account scientometric parameters such as the numbers of publications and citations, student/faculty ratio, percentage of international students, Nobel and other prizes received, number of highly cited researchers, number of papers, articles published in the Science and Nature journals, the h-index for both researchers and institutions, and web visibility. However, similar to the *no free lunch* (NFL) theorem [9], these rankings contain conceptual and methodological problems due to the underpinning bibliometric methods [10]. Despite the criticism, they continue to be one of several tools

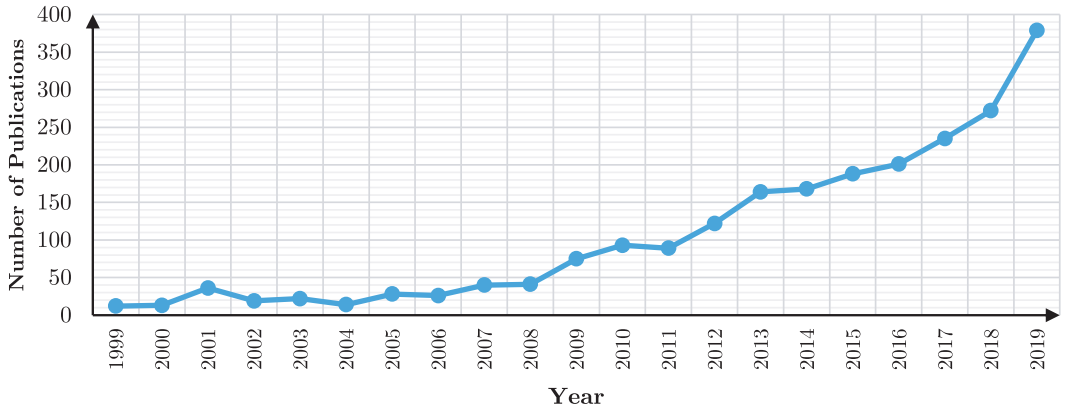

FIGURE 1. Number of publications (indexed by Scopus) related to scientometric studies reported since January 1999 to December 2019.

used by students, parents, and researchers for various purposes [4], [5]. As explained earlier, rankings remain widely used for performance appraisals and institutional evaluations since rankings are 'tangible' (e.g., easily countable rather than what it counts).

Measuring university quality has been widely studied through the analysis of university rankings. For example, [11] studied the existence of underlying entity profiles, characterized by institutions generally from the US that enjoy a high reputation on the two most influential global university rankings in the world: the Academic Ranking of World Universities (ARWU) of the University of Shanghai Jiao Tong, and the Times Higher Education (THE) ranking. Next, [12] analyzed eight different university ranking systems to measure university quality through a Bayesian hierarchical latent trait model. The author claimed that the difference between universities ranked from 50^{th} to 100^{th} and from 100^{th} to 250^{th} is insignificant. In addition, the author commented that three of the six analyzed international ranking systems are biased toward the universities in their home country. Afterward, [13] measured the academic reputation through citation networks via PageRank. The authors found some institutional cross-citation networks in the Business, Finance, and Information Science & Library Science WoS categories, which are shown using graphs. In addition, [14] studied the ARWU and THE university rankings to determine the similarities and differences in terms of their ranking criteria, main indicators, and modeling choices. The authors revealed that both analyzed university rankings have variability in the top-ranked universities across different indexes. Reference [14] further commented that some universities having a top ranking in one leading indicator are not even ranked in the rankings of another leading indicator. A recent paper [15] analyzed 33 potential indicators for the Australian higher education sector. The authors claimed that higher education equity might be better understood through the complementary use of a broader range of indicators than a single ranking index.

The articles mentioned above and others discussed further on in this article focused on conducting correlation analysis among different university rankings, used only the features provided by the analyzed rankings, and analyzed teaching indicators. However, they leave room for creating scientometric studies (analyzing research indicators), obtaining features from other benchmark databases, and using several classifiers to corroborate the obtained results and propose new visualizations based on the patterns. It is worth mentioning that contrast patterns have not been used to analyze world university rankings from a scientometric point of view. Contrast pattern-based classification, an essential topic in supervised classification, provides classification results that can be explained (similar to a human expert) [16]. In recent times, there has been growing interest in the creation of models based on eXplainable Artificial Intelligence (XAI) [17], [18], and contrast pattern-based classification is one such XAI model. Thus, in this article, we propose using a contrast pattern-based model to identify and provide patterns that describe the top universities in the rankings and that can subsequently serve as a guide for other universities to formulate their research strategies (e.g., identify areas of improvement, such as international collaboration and exposure).

The main contributions of this article are as follows:

- This is the first contrast pattern-based scientometric study comparing the top 100 world universities (according to QS ranking), with the remaining universities ranked from 101 to 200. The choice of these two classes is because many universities aspire to be in the top 100 in their vision to be world class [19]–[22].
- The paper compiled a database that contains 34 features, which can be used to describe relevant research outputs from these top 200 ranked world universities.
- Using 18 supervised classifiers and our collected database, the experiments show that the top 100 world university ranking is separable from the remaining 100 universities with an average accuracy of 71%.

- This is the first paper providing a set of contrast patterns describing the top 100 world university ranking and another set describing universities ranked from 101 to 200 in the world university ranking. The extracted patterns can potentially help decision-makers take action(s) to improve their universities' research policies.
- This is also the first paper demonstrating the use of visualization on the extracted scientometric indicators (i.e., using the information provided by the extracted patterns). By using these visualizations and the extracted patterns, experts in the application area can interact and obtain relevant information for decision-making.

The remaining of this article is structured as follows. Section II presents the relevant background materials on the world university rankings and contrast pattern-based classification. Section III presents the related scientometric literature on the world university rankings. Section IV describes our experimental setup in terms of the data acquisition, database description, and tested classifiers. Section V presents and discusses the findings from this scientometric study. Finally, Section VI concludes this article.

II. BACKGROUND OF THE STUDY

Since this article analyses a world university ranking from a scientometric point of view based on contrast patterns, this section reviews several prominent world university rankings (Section II-A) and the concepts related to pattern-based classification (Section II-B).

A. WORLD UNIVERSITY RANKINGS

The globalization of the higher education sector has contributed to an increase in competition between universities, such as that for students (including international students who are often called ''cash cows,'' since they generally pay significantly higher tuition and fees.[3] Hence, it is not surprising that universities are finding ways to increase their attractiveness to prospective students (and researchers alike), such as by placing higher in university rankings [1]. This is partly evidenced by the observation that a large number of top universities have dedicated webpages on their websites listing their positions in the various university rankings (e.g., Australian National University,[4] Nanyang Technological University in Singapore,[5] and Massachusetts Institute of Technology in the U.S.[6]). Having a better position in the world university rankings can also increase the attractiveness of the universities to funding agencies, prospective researchers and faculty members, and other universities seeking to collaborate on both students and research exchange, which consequently increase their research performance [4], [5].

Next, each of the prominent university rankings, which have been used in several scientometric studies [4], [5], [10], is described. Then, a comparative summary of these rankings is provided.

The Academic Ranking of World Universities (ARWU[7]) has been published annually since 2003 by the Institute of Higher Education, Shanghai Jiao Tong University. ARWU includes a ranking of global universities that takes into account the alumni and staff who are the recipients of Nobel and/or similar prestigious prizes, the number of articles published in *Nature* and *Science*, the number of highly cited researchers selected by Clarivate Analytics, the number of articles indexed in the Science Citation Index - Expanded and Social Sciences Citation Index, and the per capita performance of a university [1], [24].

The Ranking Web of World Universities or Webometrics Ranking (WR[8]) has been published since 2004 by the Cybermetrics Lab [24], [25], a research group of the Spanish National Research Council (CSIC). WR uses web data extracted from search engines, including documents in rich formats, the number of webpages, the papers indexed by Google Scholar, and the number of external links as a measure of link visibility or impact [1].

The Performance Ranking of Scientific Papers for World Universities has been published by the Higher Education Evaluation and Accreditation Council of Taiwan (HEEACT[9]) since 2007. HEEACT is based on the number of publications and citations according to the Thomson ISI citation databases (Science Citation Index, Social Science Citation Index, and Essential Science Indicators), particularly focusing on recent publications [1].

The Quacquarelli Symonds (QS) World University Rankings of the world's top universities (QS[10]) produced by Quacquarelli Symonds has been published annually since 2004. The QS rankings use six distinct indicators, namely, the following: academic prestige according to an extensive survey (for example, in 2008, QS used more than 6,000 respondents), the results from an employer survey, the student-faculty ratio, citations per capita according to the Elsevier Scopus database, and the proportions of international professors and international students [1].

The Times Higher Education (THE[11]) World University Rankings has produced and published annually since 2004 and in partnership with Thomson Reuters since 2010. THE is designed to provide detailed performance information across all of the core areas of university activity, as well as allow comparisons and benchmarking against other institutions across regions, subjects, and other key criteria [7]. THE uses the following indicators: the learning environment,

[3] In Australia, for example, it was reported by Universities Australia (the peak body for the higher education sector in the country) that ''International students injected $31.9 billion into Australia's economy'' [23].

[4] https://services.anu.edu.au/planning-governance/performance-measurement/world-rankings

[5] http://www.ntu.edu.sg/AboutNTU/CorporateInfo/Pages/university-rankings.aspx

[6] http://news.mit.edu/topic/rankings

[7] www.shanghairanking.com

[8] www.webometrics.info/en

[9] www.heeact.edu.tw/en/

[10] www.topuniversities.com

[11] https://www.timeshighereducation.com/world-university-rankings

TABLE 1. Principal indicators for ranking universities. All indicators were extracted from each ranking web page (accessed on September 13, 2019) and by consulting the papers of [1] and [7].

Indicator	ARWU	QS	WR	HEEACT	THE
Universities analyzed	1,800+	4,388	20,000+	3,500	1,400+
Universities in the final ranking	1,000	1,000	11,997	500	1,000
Quality of education	Alumni Nobel & Field (10%)	Students/staff ratio (20%)	-	-	Learning environment (30%)
Internationalization	-	Students (5%) and Staff (5%)	-	-	Students (2.5%), staff (2.5%), and collaboration (2.5%)
Size	Size of institution (10%)	-	Web size (20%)	-	-
Research output	Nature & Science (20%), SCI & SSCI (20%)	-	Rich files (15%) and Google Scholar (15%)	Research productivity (20%)	Survey (18%), research income(6%), and productivity (6%)
Impact	Highly cited researchers (20%)	Citations in Scopus (20%)	visibility based on links (50%)	Research impact (30%)	Citations (30%) and Knowledge transfer (2.5%)
Prestige	Staff Nobel & Field (20%)	Academic reputation (40%) and Reputation employers (10%)	-	Prestige (50%)	-

number of studies, research influence, industry income, and international outlook.

Table 1 summarizes the key indicators used in the ranking of universities by ARWU, WR, HEEACT, QS, and THE. For each university ranking, the table shows the number of universities analyzed; the number of universities in the final ranking; and how they measure the quality of education, internationalization, size,[12] research output, impact, and the prestige of the institution according to the international community. Additionally, the table shows the weight for each indicator used in the formula of each analyzed university ranking.

There are a number of different university rankings, each with its own set of criteria. Hence, it is challenging to say which university ranking is better than the other(s) because they all have drawbacks and advantages. Many universities select at least one of these rankings when publicizing or formulating their research strategies. In this study, we use the QS ranking because this ranking is widely used in US and Mexican universities, where the authors work.

B. CONTRAST PATTERN MINING

Currently, there is a trend of moving away from black-box artificial intelligence models towards eXplainable Artificial Intelligence (XAI) models, particularly for critical problems such as healthcare, criminal justice, finance, and the military [17]. The main reason is that XAI models can provide both good classification results and an explanation of the model in a language close to experts in the application area [16], [26].

A family of algorithms following the XAI approach is the contrast pattern-based classification, which allows for obtaining both accurate classification results and understandable models [16].

A *pattern* can be represented by a conjunction of relational statements (a.k.a. *items*), each with the form: $[f_i \ \# \ v_j]$, where v_j is a value in the domain of feature f_i, and # is a relational operator from the set $\{\in, \notin, =, \neq, \leq, >\}$ [16], [26]. For example, $[Number\_of\_articles \in [1000, 4000]] \wedge [Number\_of\_authors > 6000] \wedge [Number\_of\_articles\_in\_Q4 \leq 50] \wedge [Inter\_collab = \text{“}Yes\text{”}]$ could be a pattern describing a collection of universities in the top 50 of the QS ranking. In this way, a pattern is labeled as a *contrast pattern* when it appears significantly in a class regarding the other classes. Similarly, a pattern is labeled as a *pure pattern* (a.k.a. a jumping pattern) when it appears significantly in only one class of the problem [16], [26].

Contrast-pattern classification contains three stages: *mining* for extracting patterns, *filtering* for obtaining a small collection of high-quality patterns, and *classification* for combining the information of the patterns to classify query objects. In this article, only the mining stage for extracting contrast patterns [16], [26], [27] will be used; consequently, the next paragraphs focus on the mining stage.

The mining stage aims to find patterns from a data collection by using an exhaustive-search approach or using decision tree-based algorithms [18]. Regarding these options, the best option is to use algorithms based on decision trees because the local discretization performed by decision trees with numeric features has been proved to achieve better results than using the a priori global discretization used by the exhaustive-search approach. Additionally, decision trees contain a small proportion of candidate features, even in longer tree paths, which reduces the search space of potential patterns significantly and generates a small collection of high-quality patterns. Additionally, those contrast pattern mining algorithms based on decision trees can handle missing values by introducing a penalty factor in the measure for evaluating candidate splits [16], [27].

There are several contrast pattern mining algorithms based on decision trees [16], [26]–[28], but HRFm [16] allows one to obtain good classification results in both balance and imbalanced problems. HRFm builds several random decision trees by using the Hellinger distance [29] for evaluating the

[12]Size is computed by using the total scores of all remaining indicators divided by the number of full-time equivalent academic staff.

split candidates. HRFm was used in [16] for extracting contrast patterns where it allows obtaining a good collection of high-quality patterns. That is why this study will use HRFm for extracting contrast patterns from our collected database.

III. RELATED WORK

A scientometric study aims to quantify aspects of science as a discipline [30]. Consequently, several scientometric studies have been published that show the scientific activities of a specific area of science [31]. However, since there are several published papers conducting scientometric studies, this section focuses on those papers related to university ranking-based scientometric studies emphasizing research works.

One of the pioneer papers about the methodological problems presented by the Academic Ranking of World Universities (ARWU) of the Shanghai Jiao Tong University was presented by [32]. The authors claim that the quality of higher education institutions is not easy to measure or compare on an international basis. However, they said that a way to construct a reliable ranking of the world's universities is by using a comparative display of research performance because the research output is perfectly measured by using public databases. Nevertheless, the authors state that the ranking issued by ARWU in 2004 only considered those articles labeled as an *original article*, which is unfair because those articles labeled as *communication* or books are also important sources of original research. Additionally, the authors commented that the university rankings affect those universities having more than one commonly-used name, such as the *Université de Paris 6* that is also called the *Université Pierre & Marie Curie*. A similar problem is presented by those universities in non-English speaking countries, which often have different names for the same university or problems related to the variations in the translation. The main drawback of the paper presented by [32] is that the research is based only on the ARWU ranking.

In [24], the authors presented an idea for creating a ranking of world universities. The authors took into account the number of published web pages, the number of rich files, the number of articles gathered from the Google Scholar database, and the total number of external links to create a measure for ranking 13,043 universities from 191 countries. From the 10.2 million Google Scholar records and 39 million Google rich files, the author concluded that the number and positions of the EEUU universities are far more significant and better than their European counterparts. The best-ranked universities were Cambridge, Oxford, Harvard, MIT, Stanford, and Berkeley. The main drawback of the proposal presented by [24] is that those universities having open access files obtained better rankings than others having private files. Additionally, the authors did not take into account the number of citations of each author analyzed on Google Scholar. Furthermore, the author did not use any data mining algorithms to compare their proposal against the other state-of-the-art proposals.

Then, in [1], the authors proposed a study that compares university rankings. The authors compared the following five rankings:

- The rankings of the Shanghai Jiao Tong University (ARWU) for the years 2005-2008.
- The QS World University Rankings of the world's top universities (QS) for the years 2005-2008.
- Web Ranking of World Universities (WR) by the Cybermetrics Lab at CSIC for the years 2006-2008.
- The rankings of the Higher Education and Accreditation Council of Taiwan (HEEACT) for the years 2007-2008.
- The ranking of the Centre for Science and Technology Studies at Leiden University (CWTS).

From these rankings, the authors analyzed six essential indicators, such as the quality of education, internationalization, size, research output, impact, and prestige. From their study, the authors concluded that CWTS has a very advanced merging policy and excludes organizations with low publication performance. QS excludes biomedicine-only institutions from its global ranking. The authors mentioned that ARWU and QS are the most prominent rankings and when rankings from different years were used to compare their similarities, the results were the following: (i) The similarities between the ARWU rankings for the different years are very high while (ii) the similarities between the different QS rankings are much lower. The main drawbacks of the study presented by [1] are that the authors only used a correlation method to compare all the analyzed rankings, and they did not use some data mining tools to provide patterns indicating the advantages of each analyzed ranking.

In [33], the authors explored the biasing effects in reputation scores for the world university rankings. The authors claimed that ARWU could influence QS in the first years since 13 of the top 15 schools in the ARWU in 2003 were from the U.S. Additionally, the authors commented that those rankings using a survey to evaluate some indicators could be influencing the survey results. Consequently, the authors claim that one solution would be to ask respondents to rate the universities for which they have in-depth knowledge. Unfortunately, this will likely generate a conflict of interest because people always have subjective favoritism. The main drawback of their proposal is that the authors have limited the results only to those comparing the two rankings of ARWU and QS from a statistical point of view, giving no value to QS ranking because it uses survey results.

Then, in [2], the authors proposed a study of French and German universities of excellence to improve the quality assessment of the composite indicators in university rankings. The main findings of [2] were a new general methodology for building robust rankings based on simulation techniques and proving the behavior of that methodology by benchmarking it against some European universities. The authors used random simulations, intending to mitigate the potential bias in the selection of weights. Consequently, they proposed a way to rank universities according to a plurality of possible

scenarios. The main drawbacks of their proposal are the following: (i) it is a study limited to French and German universities and comparison with other countries is missing and (ii) their proposal uses a set of indicators that are different from those usually used by the top world university rankings.

In [3], the authors proposed to study the ranking of the research output of Spanish universities based on the multi-dimensional prestige of influential fields. As a consequence, the authors proposed a new approach to the ranking of the research production of universities over scientific fields based on a multivariate performance indicator space, which integrates both quantitative and qualitative dimensions. The authors used 56 universities from Spain and compared them during the 2006-2010 period. Additionally, they used the following six indicators:

1) Number of citable papers (articles, reviews, notes, or letters) published in scientific journals.
2) Number of citations received by all citable papers.
3) *H*-Index.
4) Ratio of papers published in journals in the top JCR quartile.
5) Average number of citations received by all citable papers.
6) Ratio of papers that belong to the top 10% most cited.

Based on their results, the authors claim that the top 10 universities from Spain during the 2006-2010 period were the following: Barcelona, Autónoma de Barcelona, Autónoma de Madrid, Valencia, Complutense de Madrid, Granada, Santiago de Compostela, Zaragoza, Politécnica de Valencia, and Rovira i Virgili. The main limitations of the results presented by [3] are the following: (i) their results only took into account universities from Spain (although they considered citations from all countries); (ii) they did not provide patterns for describing the behavior of the best universities from Spain; and (iii) they did not take into account postgraduate students, which are very important for the research output of a university.

In [6], the author analyzed the South African universities in world rankings. The author analyzed 23 South African universities by using ARWU, QS, WR, and HEEACT; where five of South Africa's 23 universities are ranked in one or more of these four prominent world rankings. After the analyses, the author noted that South Africa's universities have low values on all the research indicators. The author concluded that South African universities need to increase publications, citations, faculty-student ratios, and proportions of postgraduate students, international students, and international staff to rise in the rankings. The main drawbacks of the results presented by [6] are that (i) it is limited to South African universities and (ii) the author did not provide any pattern that contrasts the universities.

In [34], the authors explored a method for evaluating the comprehensive competitiveness of American universities in Bridge Engineering. The authors used the Essential Science Indicators (ESI), SCI, and EI databases to obtain the research's data and QS to obtain the ranking. The authors collected all those papers related to the Bridge Engineering field from the ESI, SCI, and EI databases. The authors claim that the University of Illinois-Urbana-Champaign and Georgia Institute of Technology are the most competitive institutions in the field of Bridge Engineering. The main drawback of the results presented by [34] is that they limit their findings to the Bridge Engineering field, which is very specific.

In [7], the author examined the methodological approach underpinning the construction of university rankings, which mainly involves calculating a composite index from an array of data. The author analyzed the ARWU, QS, and THE rankings by using a set of the top- 100 universities ranked in each aforementioned ranking. Then, the author used a strategy to create six clusterings of universities, which were analyzed by using the indicator of each selected ranking. From the results, the author concludes that the composite indexes do not adequately reflect the information contained in the collected data and, in this way, the rankings suggest to people that the values of the measured indicators are potentially significant when there is a slight difference among them from a statistical point of view. The main drawback of the results obtained by [7] is that the author did not provide a guide or patterns to contrast the universities.

In [35], the authors created a survey of university ranking systems for assessing research performance and academic quality indicators. The authors analyzed 13 ranking systems from which six of them are 100% focused on research performance. From their review, the authors concluded that current indicators are inadequate for assessing research outcomes accurately and should be supplemented and expanded using standardized criteria. Additionally, the authors claim that it is important to accentuate quality over quantity to support research performance improvement initiatives and outcomes. The authors commented that the most useful feedback for research improvement is by using a combined approach from the following rankings: Leiden, Thomson Reuters' Most Innovative Universities, and the SCImago ranking systems. Although the authors created an extensive review and provided excellent results, their main drawback is that they do not provide the aforementioned combined approach.

In [36], the authors studied seven university rankings through Principal Component Analysis [37], concluding that these rankings are similar in terms of that the number of publications and citations can explain the position of a university. The authors claimed that ''size matters when explaining institutional league tables.'' There are several ''big'' universities which are not in the top ranks proving that the size is not enough for a university because it also needs ''efficiency''; i.e., it requires that most of its professors publish one or more papers a year (we elaborate on this statement in the conclusions section).

Our review concluded (see Table 2) that there is a continuing interest in improving and analyzing the different proposed world university rankings. Additionally, there is no scientometric study based on the top world university rankings

TABLE 2. Summary of advantages and disadvantages of previous works about scientometric studies of university rankings. The meaning of the acronyms is the following: ARWU: Academic Ranking of World Universities, QS: Quacquarelli Symonds, WR: Web Ranking of World Universities, HEEACT: Higher Education and Accreditation Council of Taiwan, CWTS: Centre for Science and Technology Studies at Leiden University, and THE: Times Higher Education.

Reference	Advantages	Disadvantages
Liu and Cheng [32]	Some drawbacks identified in the ARWU ranking might also apply for other rankings; e.g., the ranking does not unify the multiple names that a university could have in its papers.	The research was based only on the ARWU ranking. It did not include a scientometric analysis of the universities and their respective rankings.
Aguillo et al. [24]	It proposed a ranking based on millions of public records.	The proposed ranking prioritizes universities with a higher number of open access articles. The authors did not compare their proposal with other state-of-the-art rankings using data mining algorithms.
Aguillo et al. [1]	It compared five university rankings (ARWU, QS, WR, HEEACT, and CWTS).	The comparison is based only on correlations. They did not use a data mining algorithm to discover the patterns of each studied ranking.
Bowman and Bastedo [33]	It explored the biasing effects in reputation scores for the world university rankings.	The statistical comparison of ARWU and QS did not include mining patterns to make scientometrics recommendations.
Benito and Romera [2]	It proposed a new general methodology for building robust rankings based on simulation techniques.	It is a study limited to French and German universities. This study did not extract patterns for contrasting the rankings or the universities.
García et al. [3]	It described the universities based on a tuple of categorical and numerical indicators.	It worked only with Spanish universities. This ranking did not take into account the postgraduate students. The authors did not provide patterns for describing the behavior of the best universities.
Matthews [6]	It compared universities based on four rankings (ARWU, QS, WR, and HEEACT).	It is limited to South African universities. The authors did not provide any comparison among universities in terms of contrast patterns.
Wang et al. [34]	It collected data from different databases.	It is limited to the Bridge Engineering field. There was not a pattern analysis for describing the best universities.
Johnes [7]	It analyzed the universities based on three rankings (ARWU, QS, and THE).	There was not a pattern analysis for describing the best universities.
Vernon et al. [35]	The authors analyzed 13 ranking systems from which six of them are 100% focused on research performance. The authors proposed to combine three rankings to obtain the most useful feedback for research improvement.	The authors did not mention how to combine the rankings. The authors did not extract patterns for contrasting the universities' rankings.
Robinson et al. [36]	They performed a correlation analysis to identify the similarities among the results of seven different rankings.	The authors did not extract measurable contrast patterns to help the universities to compare with other universities in different rankings.

that provides a set of contrast patterns characterizing the top-ranked universities. We selected Qs ranking for our study because, in Mexico, this ranking is used as a reference for comparing among the universities and for attracting the best students and professors. Based on all these conclusions, the next sections propose using SciVal[13] and the QS ranking to extract contrast patterns and, as a result, provide a guide to universities on improving their position in the QS ranking from a scientometrics approach.

IV. METHODOLOGY

Our methodology consists of the following steps: Section IV-A data acquisition describes all the tools and sources used for extracting the data; Section IV-B describes all the extracted features and gives a correlation study about these features; and Section IV-C shows the selected classifiers, their parameter values, and the source of the classifiers.

A. DATA ACQUISITION

This study uses SciVal[14] to extracting all the data. SciVal provides several metrics for assessing the research performance of over 14,000 research institutions and their associated researchers from 230 nations worldwide [38]. SciVal was selected because it is based on Scopus and allows for obtaining several metrics related to the authors and institutions, such as collaboration impact, field-weighted citation impact (FWCI), publications in top journal percentiles, and scholarly output, which are not found in the Scopus database in an easy way [39], [40]. It is worth remarking that the publications listed or hosted in institutional repositories, Google Scholar,[15] Microsoft graph,[16] and archives (e.g., Social Science Research Network (SSRN)) are not used since these sites include nonpeer review or unpublished articles (e.g., technical reports and early drafts).

The Application Programming Interface (API) provided by Elsevier Developers,[17] which allows obtaining in each query up to 2,000 metric requests[18] per month per user, is used to acquire the data. SciVal's API also allows one to obtain data for the last five years, including retrieving the following metrics: scholarly output, publications in top journal percentiles, outputs in top citation percentiles, field-weighted citation impact (FWCI), collaboration, citation count, citations per publication (CPP), h-indices, collaboration impact, and cited publications. These metrics and other ones will be explained in the next section.

B. DATABASE DESCRIPTION

As was stated in Section III, this article aims to use SciVal and the QS ranking to extract contrast patterns describing the top-ranked universities from a scientometrics approach. To do that, we collected all information provided by SciVal for the top 200 universities based on the 2020 QS ranking.[19] Then,

[13] SciVal provides several metrics extracted from Scopus
[14] https://www.elsevier.com/solutions/scival
[15] https://scholar.google.com/
[16] https://developer.microsoft.com/en-us/graph
[17] https://dev.elsevier.com/scival_apis.html
[18] https://www.elsevier.com/__data/assets/pdf_file/0020/554114/Presentation-Deck.pdf
[19] https://www.topuniversities.com/university-rankings/world-university-rankings/2020

the QS ranking was divided into two classes: the top 100 universities belong to a class, and the remaining ones belong to another class. Additionally, the following features have been extracted:

intColYYYY: Percent of publications reporting international collaboration for the year YYYY.[20]
acaColYYYY: Percent of publications reporting collaborations from other educational institutions for the year YYYY.
pubYYYY: Number of articles published in the year YYYY.
citYYYY: Number of citations received in the year YYYY.
pubTCPYYYY: Number of publications in the top 10% of the most-cited publications in the year YYYY.
pubTJPYYYY: Number of publications in the top 10% of the most-cited journals in the year YYYY.
citPPYYYY: Ratio of citations per publication (i.e., $\frac{citYYYY}{pubYYYY}$) for the year YYYY.
authorsYYYY: Number of authors publishing in the year YYYY.
citPAYYYY: Ratio of citations per author (i.e., $\frac{citYYYY}{authorsYYYY}$) for the year YYYY.
pubPAYYYY: Ratio of publications per author (i.e., $\frac{pubYYYY}{authorsYYYY}$) for the year YYYY.
h5Index: The h-index computed over the last five years.
fwCitImpYYYY: Number of citations received by an entity's publications compared with the average number of citations received by all other similar publications in the data universe for the year YYYY. The fwCitImpYYYY value indicates if the entity's publications have been cited exactly the same, more, or less than would be expected based on the global average for similar publications. For example, a value of 2.11 means 111% more than the world average, a value of 0.87 means 13% less than the world average, and a value equal to 1 means that it was exactly as expected.

The information was collected from 2016 to 2018, taking into account the 12 previously described features. This results in a two-class database containing 34 features and 200 objects (100 objects per class). Then, contrast patterns were extracted to describe the top 100 universities and the remaining universities in the 101^{th} to 200^{th} positions. The extracted patterns will provide useful information for the universities to improve their research policies. Other features (e.g., arithmetic combinations of the previously described features) did not improve the accuracy of the tested classifiers, and they provided long patterns.

Fig. 2 shows a correlation matrix formed using the Pearson's correlation method [41] and the 34 proposed features from our collected databases. The correlation is a measure of association between two numerical variables. Pearson's correlation test returns values close to one for strong positive correlations, close to zero for non-correlated variables, and close to -1 for strong negative correlations [41].

Fig. 2 shows a high correlation among similar features from the three years (2018, 2017, and 2016). This means that the scientometric measures of the best 200-ranked universities are consistent in a three-year interval.

The features pubYYYY, citYYYY, authorsYYYY, and h5index have higher absolute negative correlations (ranging from -0.4 to -0.5) with the class value than those of the other features. Given that a label with a value of 1 was used for the best 100-ranked universities and a label with a value of 2 was used for the 101-200-ranked universities, it is inferred that the best-ranked universities have more publications (pubYYYY), citations (YYYY), and authors (authorsYYYY) than the worst-ranked universities.

The features intColYYYY, acaColYYYY, citPAYYYY, and pubPAYYYY have a lower absolute negative correlation, have no correlation or have a low positive correlation (ranging from -0.1 to 0.1) with the class value than those of the other features. Hence, international collaboration, academic collaboration, citations per author, and the publications per author have low impact positions for the universities in the best rankings.

Finally, the feature authorsYYYY has strong positive correlations with the features pubYYYY, citYYYY, and h5index but a negative correlation with pubPAYYYY, citPAYYYY, intColYYYY, acaColYYYY, and class. Hence, it is concluded that having many authors (no matter the publications per author, citations per author, international collaboration, and academic collaboration) is what most impacts the rankings of the universities because having many publications, many citations, and a high h5index comes with many authors.

The next section complements the above findings by using white- and black-box state-of-the-art classifiers [18], [42], [43].

C. SELECTED CLASSIFIERS

For our experiments, 18 supervised classifiers proposed in the literature have been selected, which follow a black-box or interpretable approach [18].

Table 3 shows the selected classifiers and the parameters used in this article. The table shows the acronym, full name, parameters used, and if the algorithm was executed using the Weka data mining tool [44] or the scikit-learn library [45] for each selected algorithm. Notice that all the classifiers were executed using the parameter values recommended by their authors.

The main idea of selecting these classifiers is to see how accurate they are on the collected database. These classification results will show the separability of the classes. Additionally, the contrast pattern miner included in the PBC4cip classifier will be used to extract several patterns describing those universities in the top 100 QS ranking and the remaining universities ranked from 101 to 200 in the QS ranking.

[20]Each feature containing YYYY represents the year taken into account when extracting the information from SciVal.

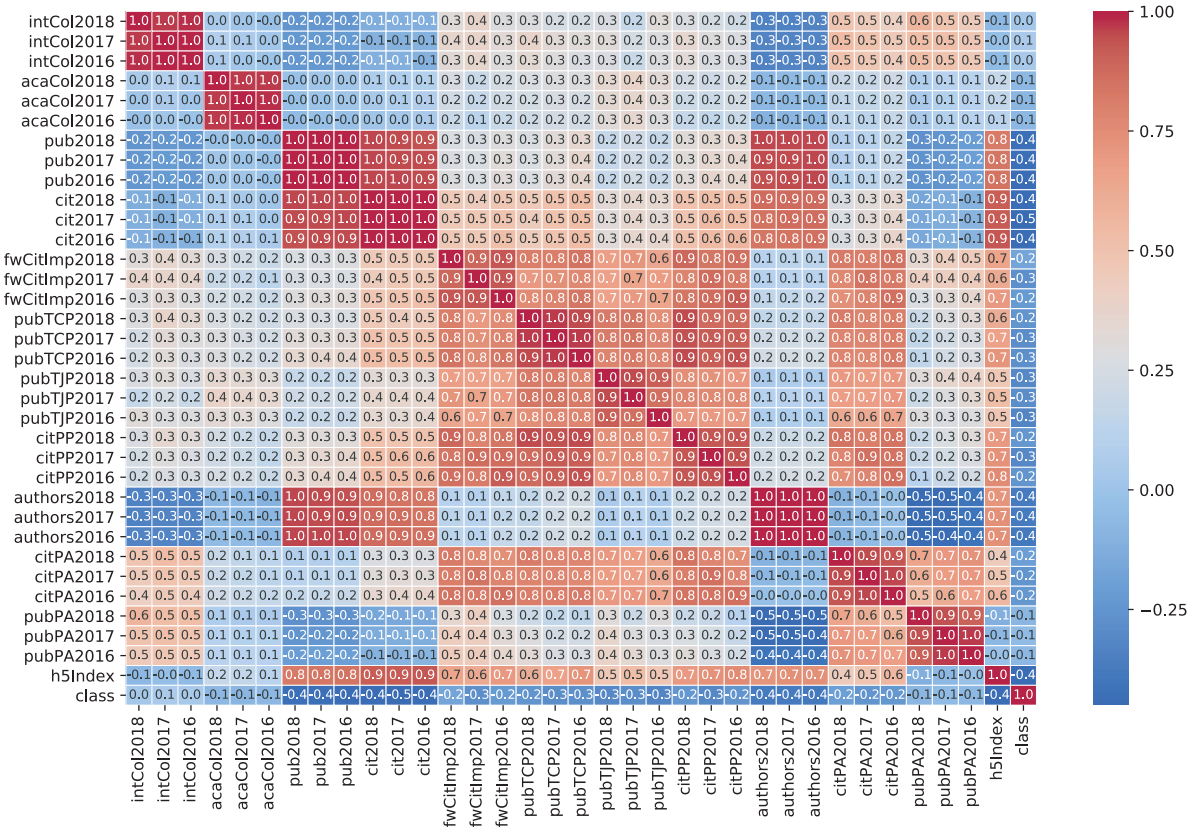

FIGURE 2. Correlation matrix for our collected database.

Since the collected database is perfectly balanced, the performance of the selected classifiers will be assessed using the accuracy, as shown in (1).

$$ACC = \frac{|WC|}{|Total|} \tag{1}$$

where $|CW|$ and $|Total|$ are the number of objects well classified and the total number of objects in the testing dataset, respectively. Finally, it is important to highlight that our collected database, as well as all supplemental materials, are provided on the following website: https://sites.google.com/site/octavioloyola/papers/SciQSRank.

V. RESULTS AND DISCUSSION

First, this section will show all the classification results obtained from the tested classifiers using our collected database (Section V-A). Then, some of the extracted patterns and our proposal for visualizing these contrast patterns in an interpretable way for experts in the application area will be analyzed (Section V-B).

A. CLASSIFICATION RESULTS

Fig. 3 shows the accuracy values obtained after testing 18 state-of-the-art classifiers. Notice that 11 of the selected classifiers (QDA, PBC4cip, Bagging, SLogistic, J48, AdaBoost, ExtraTrees, LDA, BayesNet, and GaussianNB) obtained accuracies greater than 70%. Nevertheless, five of the selected classifiers (1-NN, XGBClassifier, NaiveBayes, PART, and DClassifier) achieved accuracy values ranging from 60% to 70%. Additionally, there are two classifiers (MLP and SVM) with accuracies lower than 60%; they could be interpreted as outliers.

From Fig. 3, it is essential to highlight that PBC4cip obtains an accuracy of 73.5% on our collected database, which is an excellent result because it means that the extracted patterns can separate the problem's classes with 73.5% accuracy. In this way, the extracted patterns can be used for visualizing those universities in the top 100 QS ranking from the remaining universities ranked from 101-200 in the QS ranking in an interpretable way to experts.

Fig. 4 shows a box-plot of all the accuracies obtained by the tested classifiers on our collected database. The box-plot shows the minimum and maximum values, the median (the green line inside the box), and the first and third quartiles (top and bottom side of the box, respectively) for the accuracy. Small boxes and whiskers closer to the median indicate lower variability in the measure and, consequently, more consistent results. Accuracy values considered as outliers are shown as dots outside the whiskers. The best possible value for the accuracy is 100%, which corresponds to perfect classification.

TABLE 3. Parameter specification for the algorithms tested in our experimentation. For each parameter, detailed text can be found in the Weka data mining tool [44] or the scikit-learn library [45].

Acronym	Algorithm	Parameters	Source
QDA	Quadratic Discriminant Analysis	$priors = None, reg\_param = 0.0, store\_covariance = False, tol = 0.0001$	scikit-learn
PBC4cip	Pattern-based Classifier for Class Imbalance Problem	$weka.classifiers.trees.PBC4cip - S1 - minerRandomForestMinerWithoutFiltering - bagSizePercent100 - numFeatures - 1 - numTrees150 - builderDecisionTreeBuilder - distributionEvaluatorQuinlanGain - maxDepth - 1 - minimalObjByLeaf2 - minimalSplitGain1.0E - 30$	Weka
RND	Random Forest	$weka.classifiers.trees.RandomForest - P100 - I100 - num - slots1 - K0 - M1.0 - V0.001 - S1$	Weka
Bagging	Bagging Classifier	$weka.classifiers.meta.Bagging - P100 - S1 - num - slots1 - I10 - Wweka.classifiers.trees.REPTree - M2 - V0.001 - N3 - S1 - L - 1 - I0.0$	Weka
SLogistic	Simple Logistic	$weka.classifiers.functions.Logistic - R1.0E - 8 - M - 1 - num - decimal - places4$	Weka
J48	Decision Tree	$weka.classifiers.trees.J48 - C0.25 - M2$	Weka
AdaBoost	AdaBoost Classifier	$weka.classifiers.meta.AdaBoostM1 - P100 - S1 - I10 - Wweka.classifiers.trees.DecisionStump$	Weka
ExtractTrees	Extra-trees classifier	$n\_estimators = 100, criterion = gini, max\_depth = None, min_samples\_split = 2, min\_samples\_leaf = 2, min\_weight\_fraction\_leaf = 0.0, max\_features = log2, max\_leaf\_nodes = None, min\_impurity\_decrease = 0.0, min\_impurity\_split = None, bootstrap = False, oob\_score = False, n\_jobs = None, random\_state = 970, verbose = 0, warm\_start = False, class\_weight = None$	scikit-learn
LDA	Linear discriminant analysis	$weka.classifiers.functions.LDA - R1.0E - 6$	Weka
BayesNet	Bayesian network	$weka.classifiers.bayes.BayesNet - D - Qweka.classifiers.bayes.net.search.local.K2 - P1 - SBAYES - Eweka.classifiers.bayes.net.estimate.SimpleEstimator - A0.5$	Weka
GaussianNB	Gaussian Naive Bayes	$priors = None, var\_smoothing = 1e - 09$	scikit-learn
1-NN	k-nearest neighbors	$weka.classifiers.lazy.IBk - K1 - W0 - Aweka.core.neighboursearch.LinearNNSearch - Aweka.core.EuclideanDistance - Rfirst - last$	Weka
XGBClassifier	Extreme Gradient Boosting	$max\_depth = 9, learning\_rate = 0.1, n_estimators = 100, silent = True, objective = binary : logistic, booster = gbtree, n\_jobs = 1, nthread = None, gamma = 0, min_child\_weight = 1, max\_delta_step = 0, subsample = 1, colsample\_bytree = 1, colsample\_bylevel = 1, reg\_alpha = 0, reg\_lambda = 1, scale\_pos\_weight = 1, base\_score = 0.5, random\_state = 970, seed = None, missing = None$	scikit-learn
NaiveBayes	Naive Bayes Classifier	$weka.classifiers.bayes.NaiveBayes$	Weka
PART	Partial Decision trees	$weka.classifiers.rules.PART - M2 - C0.25 - Q1$	Weka
DClassifier	Dummy Classifier	$strategy = stratified, random\_state = 970, constant = None$	scikit-learn
MLP	Multi-layer Perceptron	$hidden\_layer\_sizes = 100, activation = relu, solver = adam, alpha = 0.0001, batch\_size = auto, learning\_rate = constant, learning\_rate\_init = 0.001, power\_t = 0.5, max\_iter = 2000, shuffle = True, random_state = 970, tol = 0.0001, verbose = False, warm\_start = False, momentum = 0.9, nesterovs\_momentum = True, early\_stopping = False, validation\_fraction = 0.1, beta\_1 = 0.9, beta\_2 = 0.999, epsilon = 1e - 08, n\_iter\_no\_change = 10$	scikit-learn
SVM	Support Vector Machine	$C = 1.0, kernel = rbf, degree = 3, gamma = auto, coef0 = 0.0, shrinking = True, probability = False, tol = 0.001, cache\_size = 200, class\_weight = None, verbose = False, max\_iter = -1, decision\_function\_shape = ovo, random\_state = 970$	scikit-learn

Fig. 4 shows that the median accuracy is closer to 71% for most of the tested classifiers. Additionally, it can be noted that there are two outliers (dots), which, with the support of the Fig. 3, are identified as the MLP and SVM.

From these experimental results, the conclusions are the following:

- The collected database contains features that can separate the problem's classes with an average accuracy of 71%.
- The contrast pattern miner used in PBC4cip can extract high-quality patterns, which allow one to obtain high classification results. These patterns allow one to separate those universities in the top 100 QS ranking from the remaining universities that are ranked from 101-200 in the QS ranking.
- These classification results could be improved by optimizing each tested classifier with our collected database by using a training-validation-testing setup. However, the main idea of our experimental setup is to show that several of the most popular state-of-the-art classifiers are suitable for obtaining an average accuracy of 71% using our collected database without optimizing parameters.

As was stated before, this study aims to analyze the QS ranking from a scientometric point of view using contrast pattern mining. Hence, the next section will be focused on analyzing the most prominent extracted patterns in this study.

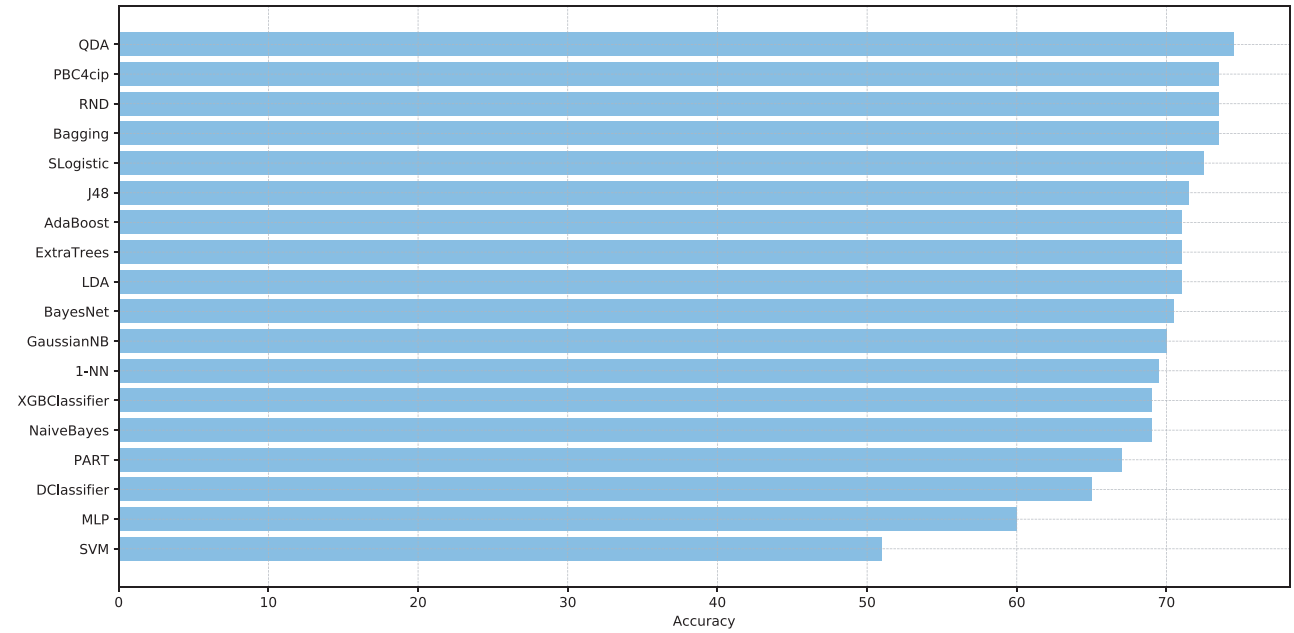

FIGURE 3. A bar chart showing the accuracy obtained by each tested classifier.

TABLE 4. Some of the most representative extracted contrast patterns.

ID	Pattern	Support by class	
		Top-100	101-200
CP_1	$authors2018 > 6722$	0.61	0.18
CP_2	$authors2016 > 6535$	0.59	0.16
CP_3	$authors2017 > 7050$	0.57	0.14
CP_4	$cit2016 > 107650$	0.54	0.06
CP_5	$cit2018 > 35720$	0.52	0.07
CP_6	$cit2017 > 75794$	0.51	0.05
CP_7	$authors2016 > 6347 \wedge cit2017 > 75225$	0.50	0.05
CP_8	$pubPA2018 > 0.99 \wedge cit2018 > 35215$	0.45	0.03
CP_9	$h5Index > 167 \wedge pub2018 > 7036$	0.45	0.04
CP_{10}	$authors2018 > 6644 \wedge h5Index > 167.50 \wedge pubTJP2018 > 35.50$	0.45	0.04
CP_{11}	$authors2018 > 6644 \wedge citPA2016 > 14.79 \wedge intCol2017 > 31.85$	0.44	0.03
CP_{12}	$authors2018 > 6644 \wedge citPA2016 > 14.79 \wedge intCol2016 > 31.10$	0.44	0.03
CP_{13}	$authors2018 > 6644 \wedge citPA2016 > 14.79 \wedge pub2018 > 7036$	0.44	0.03
CP_{14}	$cit2016 > 1173780 \wedge citPA2016 > 14.79 \wedge citPP2017 > 8.65 \wedge pubPA2016 > 0.95$	0.40	0.00
CP_{15}	$authors2017 > 6552 \wedge h5Index > 167.50 \wedge citPA2017 > 9.68 \wedge intCol2018 > 33.15$	0.36	0.00
CP_{16}	$h5Index > 172.50 \wedge pub2016 > 6489.00 \wedge citPP2017 > 8.75 \wedge pubPA2016 > 0.95$	0.36	0.00
CP_{17}	$authors2018 > 6722 \wedge citPA2016 > 14.79 \wedge pubPA2016 > 0.95 \wedge citPP2017 > 8.65 \wedge h5Index > 165$	0.40	0.00
CP_{18}	$authors2018 > 6722 \wedge citPA2016 > 14.79 \wedge pubPA2016 > 0.95 \wedge citPA2017 > 9.07 \wedge h5Index > 165$	0.40	0.00
CP_{19}	$authors2018 > 6644 \wedge citPA2016 > 14.79 \wedge pubPA2016 > 0.95 \wedge citPA2017 > 9.07 \wedge cit2016 > 117378$	0.40	0.00
CP_{20}	$citPA2016 > 14.79 \wedge cit2017 > 76797 \wedge citPA2017 > 9.07 \wedge pubPA2016 > 0.95 \wedge cit2016 > 117378$	0.40	0.00

B. ANALYZING THE EXTRACTED PATTERNS

This section shows a subset of all the patterns extracted using the contrast pattern mining algorithm included in PBC4cip.[21] To complement the mathematical representation of the contrast patterns, we visualize the data according to the patterns for better comprehension. Combining machine learning results with data visualization is becoming a popular technique currently [46]–[50] because it allows the user to see the information from different points of view.

Table 4 shows 20 of the most representative contrast patterns extracted from our collected database. For each pattern, this table shows its ID as an identifier, all items contained in the pattern, and its support by class. This table is first sorted in ascending order by the number of items and then sorted in descending order according to the support of the pattern for the Top-100 class.

Table 5 shows the top 10 most used features and items from the extracted patterns describing the top 100 universities in the QS ranking. This table shows that the most used feature is the h5index, which forms part of the most used item ($h5Index > 167$) from the extracted patterns. Notice that the top 100 universities in the QS ranking have published more than 7,400 papers in 2016; they also received more than 35,700 citations in 2018 and more than 107,600 citations in 2016. From Table 5, it can be concluded many universities in the top 100 QS ranking generate more than 7,000 papers and 35,000 citations yearly.

[21] Weka package available at https://sites.google.com/view/leocanetesi fuentes/software/multivariate-pbc4cip.

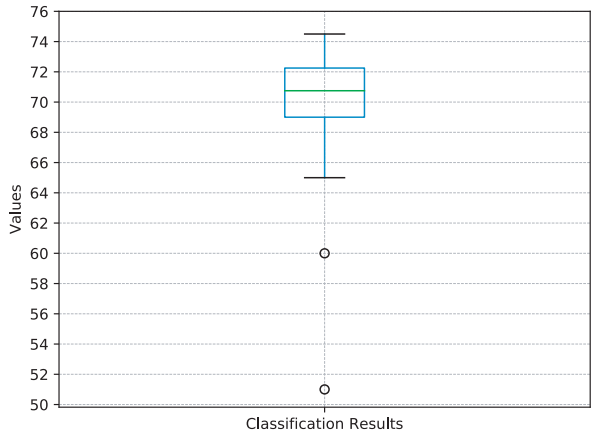

FIGURE 4. A box-plot graph showing the accuracies obtained by all the tested classifiers.

TABLE 5. List of the most used features and items from the extracted patterns describing the top 100 universities in the QS ranking.

Feature	Total	Item	Total
h5Index	455	$h5Index > 167$	233
cit2016	347	$pubPA2017 > 0.96$	181
pub2016	268	$citPA2016 > 14.79$	177
citPA2016	268	$pubPA2016 > 0.95$	166
cit2018	236	$cit2016 > 107650$	164
pubPA2017	234	$pub2016 > 7472.50$	150
cit2017	211	$cit2018 > 35720$	140
pubPA2016	177	$intCol2017 > 36.70$	131
intCol2017	171	$h5Index > 165$	130
authors2018	163	$cit2016 > 117378$	123

Table 4 shows that the set of patterns $\{CP_1, \ldots, CP_6\}$ only contains an item for each pattern, which is easy for experts in the application area to understand. Notice that each pattern in this set has at least 0.51 of support for the Top-100 class and very little support for another class.

A way to visualize those patterns having only one item is to use a box-plot (see Fig. 5). This visualization shows the area of the feature covered by the pattern in a darker color. Notice that more than 50% of the top 100 universities have more than 6, 600 authors while more than 75% of the universities in ranks 101-200 have less than 6, 600 authors.

Table 4 shows that the set of patterns $\{CP_7, \ldots, CP_9\}$ contains two items for each pattern, which are also easy for experts in the application area to understand. Notice that each pattern in this set has at least 0.45 of support for the Top-100 class and very little support for another class. It is worth noticing that these patterns include two of the most frequent features found in the patterns (i.e., citYYYY and pubYYYY). This indicates that top-ranked universities focus on achieving high numbers of citations and publications.

To visualize patterns having two items, we use a scatter plot with one color for those objects covered by the pattern and another color for those objects not covered by the pattern. Fig. 6 shows the visualization for pattern CP_7, where the circles and triangles represent those universities in the top 100 ranking and those ranked 101-200, respectively. A blue color fills those objects covered by the pattern, and a gray color fills those objects not covered by the pattern.

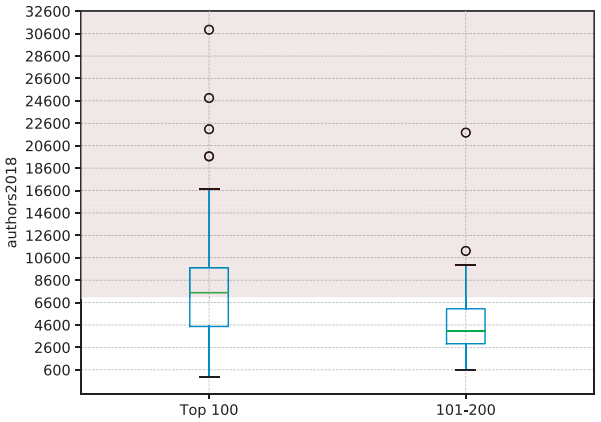

FIGURE 5. Two box-plots for visualizing pattern CP_1, which only has one item.

The analysis of pattern CP_7 from a visual perspective and taking into account its support by class shown in Table 4 shows that there is a 50% likelihood that universities in the top 100 QS ranking obtained more than 75,000 citations in 2017, and their number of authors publishing in 2016 was greater than 6,347. Notice that there is less than 5% of the universities ranked 101-200 in the QS ranking with these characteristics. Additionally, Fig. 6 shows that some universities ranked 101-200 in the QS ranking are closer to fulfilling the items described by the pattern.

Table 4 shows that the set of patterns $\{CP_{10}, \ldots, CP_{13}\}$ contains three items for each pattern. Similar to Fig. 6, Fig. 7 shows a visualization for pattern CP_{13}, where the circles and triangles represent those universities in the top 100 ranking and those ranked 101-200, respectively. A blue color fills those objects covered by the pattern, and a gray color fills those objects not covered by the pattern. Additionally, in the visualization, the third item of the pattern is defined by the color intensity within each geometric figure. Notice that Fig. 7 contains two bar legends, which represent the color intensity scale used for both those objects covered and not covered by the analyzed pattern. Notice that these patterns include one of the most frequent features found in the patterns (i.e., author2018). This indicates that top-ranked universities focus on increasing their numbers of authors publishing.

Table 4 shows that the set of patterns $\{CP_{14}, \ldots, CP_{16}\}$ contains four items for each pattern. Similar to Fig. 7, Fig. 8 shows a visualization for pattern CP_{15}, but the fourth item of the pattern is represented by the size of each geometric figure. Fig. 8 shows that those objects covered by the pattern have a borderline with those objects not covered by the pattern when the size of the geometric figure is small. Based on Table 4, it can be concluded that pattern CP_{15} is a pure pattern describing 40% of the objects belonging to the top 100 universities in the QS ranking.

Table 4 shows that the set of patterns $\{CP_{17}, \ldots, CP_{20}\}$ contains five items for each pattern. Similar to Fig. 8, Fig. 9 shows a visualization for pattern CP_{18}, but the fifth item of the pattern is represented by the size of the outer line of the geometric figure representing the problem's classes. Fig. 9

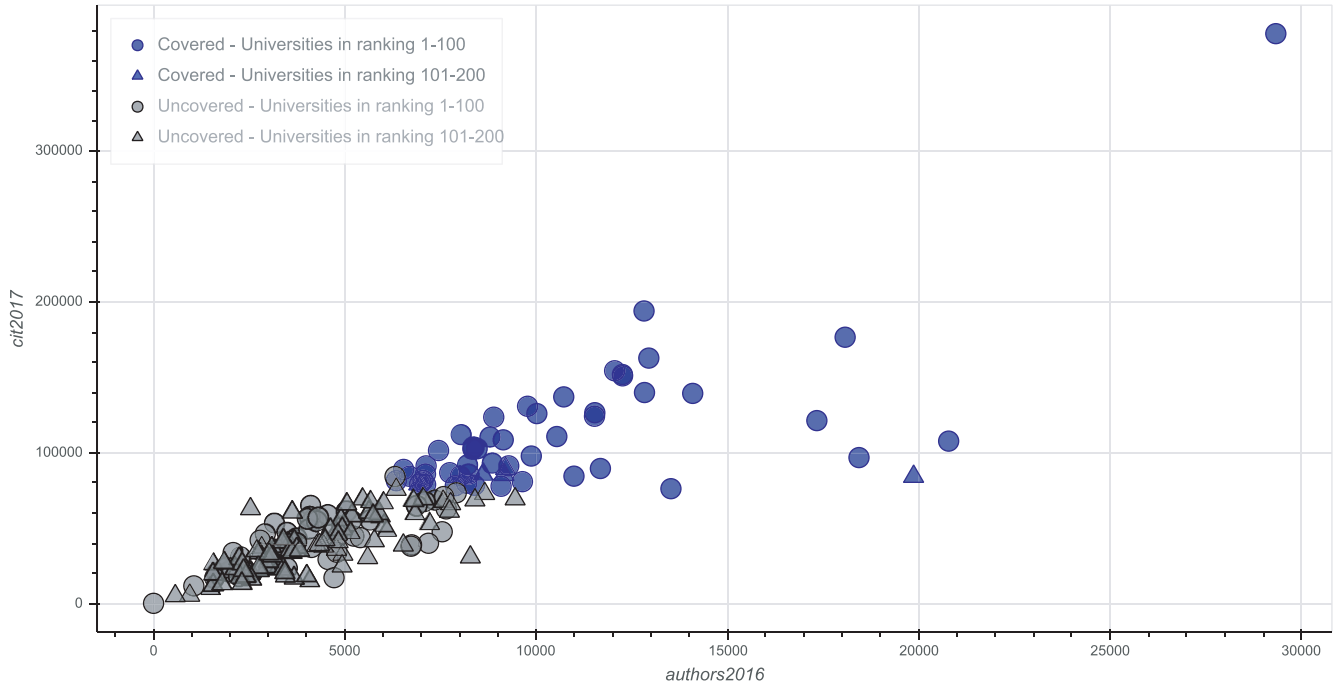

FIGURE 6. A scatter plot for visualizing pattern CP_7, which has only two items.

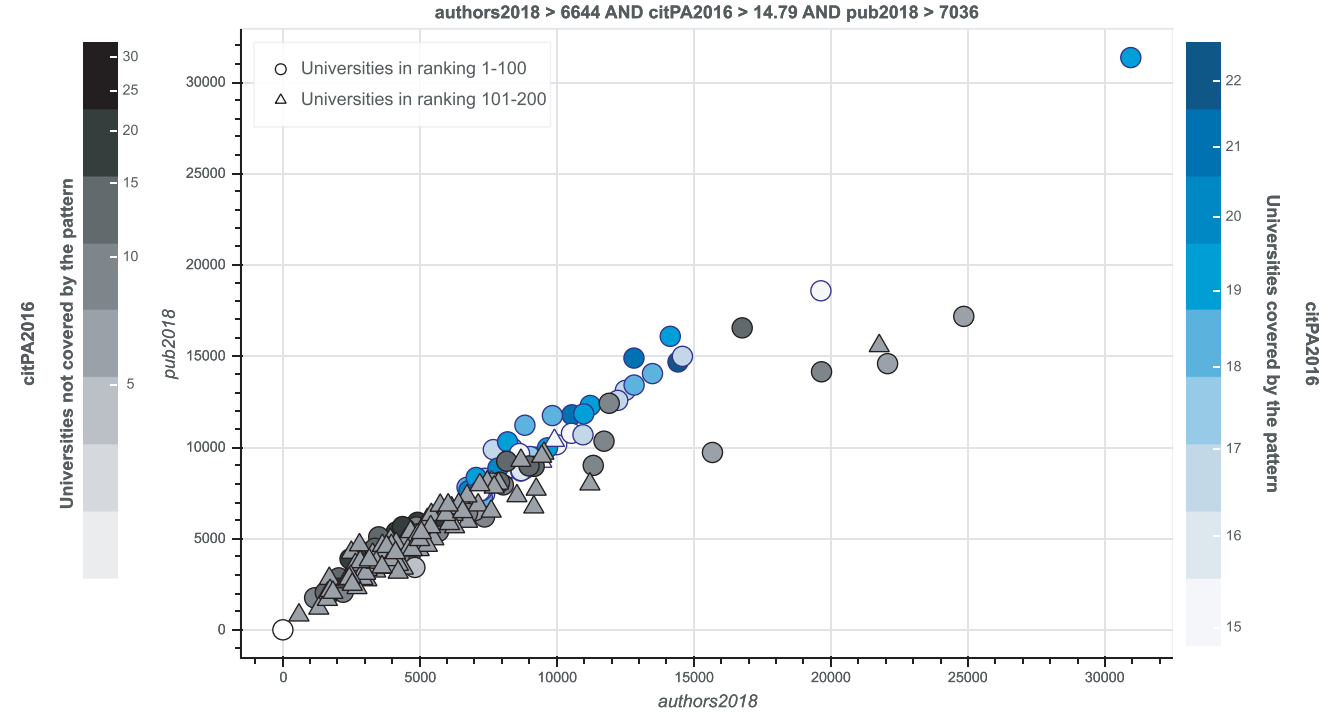

FIGURE 7. A scatter plot for visualizing the pattern CP_{13}, which has three items.

shows that those objects covered by the pattern have an overlapping zone with those objects not covered by the pattern. Based on Table 4, it can be concluded that pattern CP_{18} is a pure pattern describing 40% of the objects belonging to the top 100 universities in the QS ranking.

The following are the conclusions from the proposed visualization for contrast patterns:

- It seems to be more intuitive to understand a contrast pattern by using a visualization procedure.
- By analyzing different contrast patterns, there are universities ranked 101-200 in the QS ranking that have achieved good research results, and they could be positioned in the top 100 ranking in the coming years.

FIGURE 8. A scatter plot for visualizing the pattern CP_{15}, which has four items.

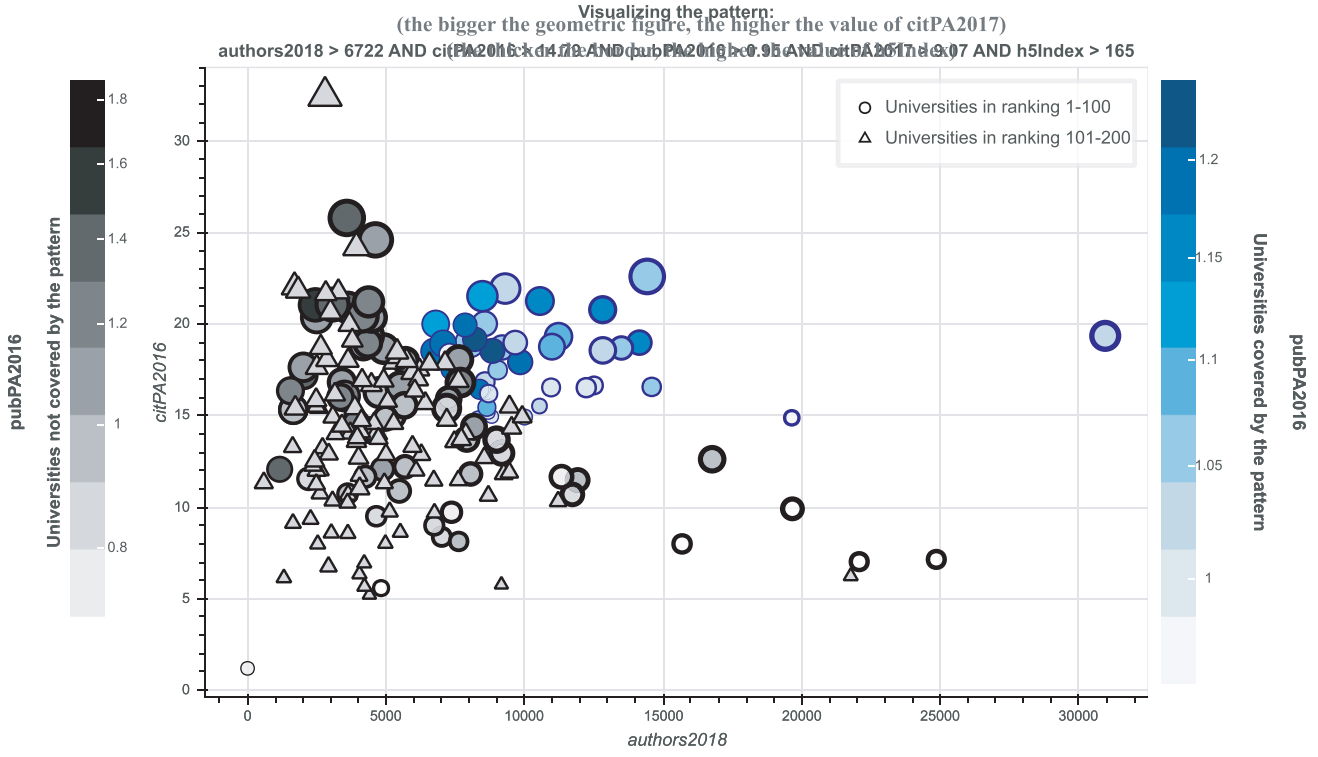

FIGURE 9. A scatter plot for visualizing the pattern CP_{18}, which has five items.

- There are universities in the top 100 QS ranking that are firmly positioned, and, consequently, these universities will continue in the top 100 for several years.
- Contrast patterns containing up to five items were visualized effectively, although the visualization procedure can be used for visualizing more than five items.

The Bokeh tool was used to create the interactive plots. Bokeh is an interactive visualization library that can be used to create interactive visualizations on modern web browsers. Bokeh provides elegant and interactive graphics over big datasets quickly and easily by using Python [51].

VI. CONCLUSION

Scientometrics can potentially contribute to the advancement of sciences, for example, by analyzing the information extracted from relevant sources to inform decision and strategy formulation. Specifically, this article introduced the first contrast pattern-based study for comparing the top 200 world-ranked universities (according to QS ranking) using a scientometric lens. This study collected a database containing 34 features that describe the relevant research outputs from these top 200 world-ranked universities. Using a set of popular state-of-the-art classifiers (taking into account white- and black-box models), we arrive at the following findings:

- The top 100 ranked universities can be clearly separated from the remaining 100 universities (i.e., those in the 101st to 200th positions) with an average accuracy of 71%.
- We extracted one set of contrast patterns describing the top 100 ranked world universities and another set describing the remaining 100 universities. The extracted patterns can be used by decision-makers to take appropriate actions to improve their universities' research policies.
- The analysis of the extracted patterns allows us to conclude that the set of top 100 ranked universities in the QS ranking is firmly positioned; consequently, these universities will most likely continue to be in the top 100 positions unless the ranking criteria significantly change. For example, these universities had more than 6,500 authors in the last studied year, published from.6 to 1.3 articles per author, achieved from 2 to 6 citations per author, and achieved an h5-index from 109 to 339. It is worth to mention that there are universities out of the top-100 rank despite that they have more than 10,000 professors (e.g., Tecnologico de Monterrey[22]). These universities could improve their research indicators if they convince most professors to publish at least one research paper a year.
- We also provided an approach for visualizing all the extracted scientometric indicators by using the information provided by the extracted patterns. By using these visualizations and the extracted patterns, experts in the application area can interact and obtain relevant information for decision-making.

Finally, as future work, our plan is to analyze other university rankings and scientometric sources to contrast these results against the ones obtained from the Scopus database. Additionally, we will use multivariate contrast patterns to describe the set of analyzed universities. Since multivariate contrast patterns contain multivariate items, which allow for linear combinations of numerical features, they are more complicated to explain and it is more difficult to create understandable visualizations from them.

[22] https://tec.mx/en/data-and-figures

ACKNOWLEDGMENT

The authors express their gratitude to the students of the postgraduate course CS5051 at the Tecnologico de Monterrey, as the class discussions contributed towards the shaping of this article.

REFERENCES

[1] I. F. Aguillo, J. Bar-Ilan, M. Levene, and J. L. Ortega, "Comparing university rankings," *Scientometrics*, vol. 85, no. 1, pp. 243–256, Oct. 2010, doi: 10.1007/s11192-010-0190-z.
[2] M. Benito and R. Romera, "Improving quality assessment of composite indicators in university rankings: A case study of French and German universities of excellence," *Scientometrics*, vol. 89, no. 1, p. 153, Aug. 2011, doi: 10.1007/s11192-011-0419-5.
[3] J. A. García, R. Rodriguez-Sánchez, J. Fdez-Valdivia, D. Torres-Salinas, and F. Herrera, "Ranking of research output of universities on the basis of the multidimensional prestige of influential fields: Spanish universities as a case of study," *Scientometrics*, vol. 93, no. 3, pp. 1081–1099, Dec. 2012, doi: 10.1007/s11192-012-0740-7.
[4] D. D. Dill and M. Soo, "Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems," *Higher Educ.*, vol. 49, no. 4, pp. 495–533, Jun. 2005, doi: 10.1007/s10734-004-1746-8.
[5] C. Eccles, "The use of university rankings in the United Kingdom," *Higher Educ. Eur.*, vol. 27, no. 4, pp. 423–432, Dec. 2002, doi: 10.1080/0379772022000071904.
[6] A. P. Matthews, "South African universities in world rankings," *Scientometrics*, vol. 92, no. 3, pp. 675–695, Sep. 2012, doi: 10.1007/s11192-011-0611-7.
[7] J. Johnes, "University rankings: What do they really show?" *Scientometrics*, vol. 115, no. 1, pp. 585–606, Apr. 2018, doi: 10.1007/s11192-018-2666-1.
[8] J. Kim and W.-J. Shim, "What do rankings measure? The U.S. news rankings and student experience at liberal arts colleges," *Rev. Higher Educ.*, vol. 42, no. 3, pp. 933–964, 2019.
[9] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.
[10] A. F. J. van Raan, "Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods," *Scientometrics*, vol. 62, no. 1, pp. 133–143, Jan. 2005, doi: 10.1007/s11192-005-0008-6.
[11] V. Safón, "What do global university rankings really measure? The search for the x factor and the x entity," *Scientometrics*, vol. 97, no. 2, pp. 223–244, Nov. 2013, doi: 10.1007/s11192-013-0986-8.
[12] C. Claassen, "Measuring university quality," *Scientometrics*, vol. 104, no. 3, pp. 793–807, Sep. 2015, doi: 10.1007/s11192-015-1584-8.
[13] F. A. Massucci and D. Docampo, "Measuring the academic reputation through citation networks via PageRank," *J. Informetrics*, vol. 13, no. 1, pp. 185–201, Feb. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S175115771830110X
[14] G. A. Olcay and M. Bulu, "Is measuring the knowledge creation of universities possible?: A review of university rankings," *Technol. Forecasting Social Change*, vol. 123, pp. 153–160, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S004016251630021X
[15] T. Pitman, D. Edwards, L.-C. Zhang, P. Koshy, and J. Mcmillan, "Constructing a ranking of higher education institutions based on equity: Is it possible or desirable?" *Higher Educ.*, vol. 80, no. 4, pp. 605–624, Jan. 2020, doi: 10.1007/s10734-019-00487-0.
[16] O. Loyola-González, M. A. Medina-Pérez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. García-Borroto, "PBC4cip: A new contrast pattern-based classifier for class imbalance problems," *Knowl.-Based Syst.*, vol. 115, pp. 100–109, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705116304002
[17] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
[18] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.
[19] E. Hazelkorn, *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence*. London, U.K.: Palgrave Macmillan, 2015.

[20] M. Benito, P. Gil, and R. Romera, "Funding, is it key for standing out in the university rankings?" *Scientometrics*, vol. 121, no. 2, pp. 771–792, Nov. 2019.

[21] E. Hazelkorn, "The impact of league tables and ranking systems on higher education decision making," *Higher Educ. Manage. Policy*, vol. 19, no. 2, pp. 1–24, Aug. 2007.

[22] E. Hazelkorn, "Reflections on a decade of global rankings: What we've learned and outstanding issues," *Eur. J. Educ.*, vol. 49, no. 1, pp. 12–28, Mar. 2014.

[23] Universities Australia. (Aug. 2, 2018). *International Students Inject $32 Billion a year Into Australia's Economy—Boosting Aussie Jobs and Wages*. Accessed: Jan. 14, 2020. [Online]. Available: https://www.universitiesaustraliaedu.au/media-item/international-students-inject-32-billion-a-year-into-australias-economy-boosting-aussie-jobs-and-wages/

[24] I. F. Aguillo, J. L. Ortega, and M. Fernández, "Webometric ranking of world universities: Introduction, methodology, and future developments," *Higher Educ. Eur.*, vol. 33, nos. 2–3, pp. 233–244, Jul. 2008, doi: 10.1080/03797720802254031.

[25] I. F. Aguillo, B. Granadino, J. L. Ortega, and J. A. Prieto, "Scientific research activity and communication measured with cybermetrics indicators," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 10, pp. 1296–1302, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20433

[26] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Cost-sensitive pattern-based classification for class imbalance problems," *IEEE Access*, vol. 7, pp. 60411–60427, 2019.

[27] M. García-Borroto, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "Finding the best diversity generation procedures for mining contrast patterns," *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4859–4866, Jul. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415001359

[28] L. Cañete-Sifuentes, R. Monroy, M. A. Medina-Pérez, O. Loyola-González, and F. V. Voronisky, "Classification based on multivariate contrast patterns," *IEEE Access*, vol. 7, pp. 55744–55762, 2019.

[29] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Mining Knowl. Discovery*, vol. 24, no. 1, pp. 136–158, Jan. 2012, doi: 10.1007/s10618-011-0222-1.

[30] W. W. Hood and C. S. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, no. 2, p. 291, Oct. 2001, doi: 10.1023/A:1017919924342.

[31] A. F. J. Van Raan, "Scientometrics: State-of-the-art," *Scientometrics*, vol. 38, no. 1, pp. 205–218, Jan. 1997, doi: 10.1007/BF02461131.

[32] N. C. Liu and Y. Cheng, "The academic ranking of world universities," *Higher Educ. Eur.*, vol. 30, no. 2, pp. 127–136, Jul. 2005, doi: 10.1080/03797720500260116.

[33] N. A. Bowman and M. N. Bastedo, "Anchoring effects in world university rankings: Exploring biases in reputation scores," *Higher Educ.*, vol. 61, no. 4, pp. 431–444, Apr. 2011, doi: 10.1007/s10734-010-9339-1.

[34] Y. Wang, R. Ma, T. Tang, X. Liu, P. Xie, J. Wang, J. Liu, H. Zhou, and S. Zhang, "The comprehensive competitiveness evaluation of American universities in bridge engineering," *Scientometrics*, vol. 91, no. 3, pp. 693–701, Jun. 2012, doi: 10.1007/s11192-012-0616-x.

[35] M. M. Vernon, E. A. Balas, and S. Momani, "Are university rankings useful to improve research? A systematic review," *PLoS ONE*, vol. 13, no. 3, pp. 1–15, Mar. 2018, doi: 10.1371/journal.pone.0193762.

[36] N. Robinson-Garcia, D. Torres-Salinas, E. Herrera-Viedma, and D. Docampo, "Mining university rankings: Publication output and citation impact as their basis," *Res. Eval.*, vol. 28, no. 3, pp. 232–240, Jul. 2019.

[37] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[38] R. Dresbeck, "SciVal," *J. Med. Library Assoc.*, vol. 103, no. 3, pp. 164–166, Jul. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511065/

[39] E. Vardell, T. Feddern-Bekcan, and M. Moore, "SciVal experts: A collaborative tool," *Med. Reference Services Quart.*, vol. 30, no. 3, pp. 283–294, Jul. 2011, doi: 10.1080/02763869.2011.603592.

[40] L. Colledge and R. Verlinde, *Scival Metrics Guidebook*. Amsterdam, The Netherlands: Elsevier, 2014.

[41] J. T. Roscoe, *Fundamental Research Statistics for the Behavioral Sciences* (International Series in Decision Processes), 2nd ed. New York, NY, USA: Holt, Rinehart & Winston, 1975.

[42] M. A. Álvarez-Carmona, E. Villatoro-Tello, M. Montes-Y-Gómez, and L. Villaseñor-Pineda, "A comparative analysis of distributional term representations for author profiling in social media," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4857–4868, May 2019.

[43] Y. Martínez-Díaz, N. Hernández, R. J. Biscay, L. Chang, H. Méndez-Vázquez, and L. Enrique Sucar, "On Fisher vector encoding of binary features for video face recognition," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 155–161, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320318300245

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[46] B. Cervantes, F. Gómez, R. Monroy, O. Loyola-González, M. A. Medina-Pérez, and J. Ramírez-Márquez, "Pattern-based and visual analytics for visitor analysis on websites," *Appl. Sci.*, vol. 9, no. 18, p. 3840, Sep. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/18/3840

[47] L. A. Trejo, V. Ferman, M. A. Medina-Pérez, F. M. A. Giacinti, R. Monroy, and J. E. Ramirez-Marquez, "DNS-ADVP: A machine learning anomaly detection and visual platform to protect top-level domain name servers against DDoS attacks," *IEEE Access*, vol. 7, pp. 116358–116369, 2019.

[48] O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez, B. Camiña, J. E. Ramírez-Márquez, and R. Monroy, "Fusing pattern discovery and visual analytics approaches in tweet propagation," *Inf. Fusion*, vol. 46, pp. 91–101, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517307716

[49] A. López-Cuevas, M. A. Medina-Pérez, R. Monroy, J. E. Ramírez-Márquez, and L. A. Trejo, "FiToViz: A visualisation approach for real-time risk situation awareness," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, pp. 372–382, Jul./Sep. 2018.

[50] M. A. Alvarez-Carmona, L. Pellegrin, M. Montes-y-Gómez, F. Sánchez-Vega, H. J. Escalante, A. P. López-Monroy, L. Villaseñor-Pineda, and E. Villatoro-Tello, "A visual approach for age and gender identification on Twitter," *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 3133–3145, 2018.

[51] Bokeh Development Team. (2019). *Bokeh: Python Library for Interactive Visualization*. [Online]. Available: https://bokeh.org/

OCTAVIO LOYOLA-GONZÁLEZ received the B.Eng. degree in informatics engineering and the M.Sc. degree in applied informatics from the University of Ciego de Ávila, in 2010 and 2012, respectively, and the Ph.D. degree in computer science from the National Institute for Astrophysics, Optics, and Electronics, Mexico, in 2017. He is currently a Researcher and a Professor with the Tecnologico de Monterrey, Campus Puebla, for undergraduate and graduate programs of computer sciences. He is also a member of the Mexican Researchers System (Rank 1). He has been involved in many research projects about pattern recognition, which have been applied to cybersecurity, biotechnology, and dactyloscopy problems. He has published several books and articles on subjects related to contrast pattern-based classification, data mining, one-class classification, masquerader detection, fingerprint recognition, and cybersecurity. He received the Best Thesis Award "José Negrete" for the Doctoral Thesis Category on Artificial Intelligence sponsored by the Mexican Society for Artificial Intelligence (SMIA), in 2018. He was the Prizewinner in the XXXI National Contest of Computer Science Thesis (ANIEI 2018) and the Best Ph.D. Thesis in the Computer Science Coordination at the National Institute of Astrophysics, Optics, and Electronics, in 2018.

MIGUEL ANGEL MEDINA-PÉREZ received the Ph.D. degree in computer science from the National Institute of Astrophysics, Optics, and Electronics, Mexico, in 2014. He is currently a Research Professor with the Tecnologico de Monterrey, Campus Estado de Mexico, where he is also a member of the GIEE-ML (Machine Learning) Research Group. He has Rank 1 in the Mexican Research System. He has published tens of articles in referenced journals, such as *Information Fusion*, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, *Pattern Recognition*, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Knowledge-Based Systems*, *Information Sciences*, and *Expert Systems with Applications*. He has extensive experience developing software to solve Pattern Recognition problems. A successful example is a fingerprint and palmprint recognition framework which has more than 1.3 million visits and 135 000 downloads. His research interests include pattern recognition, data visualization, explainable artificial intelligence, fingerprint recognition, and palmprint recognition.

RAYMUNDO ADRIÁN CORONILLA VALDEZ received the bachelor's degree in computer science from the Tecnológico de Monterrey. He is passionate about innovation and firmly believes in technology as a powerful tool to improve people's lives. He has worked on different projects such as digitalization, video game, and mobile applications development. His research interest includes generating and using tools to optimize processes and create impact.

KIM-KWANG RAYMOND CHOO (Senior Member, IEEE) holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio (UTSA). He is also a Fellow of the Australian Computer Society. In 2016, he was named as the Cybersecurity Educator of the Year–APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn). In 2015, he and his team won the Digital Forensics Research Challenge organized by Germany's University of Erlangen-Nuremberg. He was a recipient of the 2019 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence in Scalable Computing (Middle Career Researcher), the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, the Outstanding Associate Editor of 2018 for IEEE ACCESS, the British Computer Society's 2019 Wilkes Award Runner-Up, the 2019 *EURASIP Journal on Wireless Communications and Networking* (JWCN) Best Paper Award, the Korea Information Processing Society's *Journal of Information Processing Systems* (JIPS) Survey Paper Award (Gold) 2019, the IEEE Blockchain 2019 Outstanding Paper Award, the IEEE TrustCom 2018 Best Paper Award, the ESORICS 2015 Best Research Paper Award, the 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, the Fulbright Scholarship in 2009, the 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award in 2008. He is also the Co-Chair of IEEE Multimedia Communications Technical Committee's Digital Rights Management for Multimedia Interest Group.