Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

https://doi.org/10.23913/ride.v12i24.1144

Artículos científicos

Uso de la minería de datos para la caracterización de investigadores y cuerpos académicos

Use of Data Mining to Characterize Researchers and Academic Groups

Uso de mineração de dados para caracterizar pesquisadores e órgãos acadêmicos

Víctor H. Menéndez Domínguez
Universidad Autónoma de Yucatán, México
mdoming@correo.uady.mx
https://orcid.org/0000-0003-3587-1263

Jared D.T. Guerrero Sosa
Universidad Autónoma de Yucatán, México
jared.guerrero@correo.uady.mx
https://orcid.org/0000-0001-7999-9870

María Enriqueta Castellanos Bolaños
Universidad Autónoma de Yucatán, México
enriqueta.c@correo.uady.mx
https://orcid.org/0000-0001-6294-5948

José William Cervera Pérez
Universidad Autónoma de Yucatán, México
w.cervera.p@gmail.com
https://orcid.org/0000-0002-1378-3070

Resumen

En este trabajo se presentan los comportamientos y tendencias en la producción académica de profesores asociados a una universidad pública del sureste de México. Se empleó una aplicación propia que recopila y procesa la producción de los académicos publicada en Scopus. Para el análisis de la información se utilizaron técnicas de minería de datos para clasificar la producción de cuerpos académicos, la relación de académicos activos con base en el número de publicaciones recientes y su grado de certificación, los grupos definidos según los campus de la universidad, así como reglas para identificar las tendencias de los grupos de investigadores. Uno de los descubrimientos fue la correlación del número de integrantes de un grupo con su producción, y la correlación entre la producción indizada y no indizada de los académicos. Si bien el estudio fue hecho para una universidad en particular, la metodología puede ser reproducida para situaciones similares.

Palabras clave: cuerpos académicos, procesamiento de datos, producción científica, universidades.

Abstract

In this work, the behaviors, and trends in the production of researchers associated with an important public university in south-eastern Mexico is presented. A proprietary application was used that collects and processes the production of the academics published in Scopus. For the analysis of the information, data mining techniques were used to classify the production of academic bodies, the ratio of active academics based on the number of recent publications and their certification degree, the groups defined according to the university campuses, as well as rules to identify trends in researchers' groups. One of the discoveries was the correlation between the number of group members with their production, and the correlation between the indexed and non-indexed researchers' production. Although the study was done for a particular university, the methodology can be reproduced for similar situations.

Keywords: academic groups, data processing, scientific production, universities.

Resumo

Este artigo apresenta os comportamentos e tendências na produção acadêmica de professores associados a uma universidade pública no sudeste do México. Foi utilizado um aplicativo próprio que coleta e processa a produção acadêmica publicada na Scopus. Para a análise das informações, foram utilizadas técnicas de mineração de dados para classificar a produção dos corpos acadêmicos, a proporção de acadêmicos ativos com base no número de publicações recentes e seu grau de certificação, os grupos definidos de acordo com os campi universitários, bem como regras para identificar tendências em grupos de pesquisa. Uma das descobertas foi a correlação do número de membros de um grupo com sua produção, e a correlação entre a produção indexada e não indexada dos acadêmicos. Embora o estudo tenha sido feito para uma determinada universidade, a metodologia pode ser reproduzida para situações semelhantes.

Palavras-chave: órgãos acadêmicos, processamento de dados, produção científica, universidades.

Fecha Recepción: Agosto 2021 **Fecha Aceptación:** Febrero 2022

Introduction

Nowadays, a relevant factor to determine the relevance of a university lies in the scientific production that it generates (Leahey, 2016). This attribute is commonly measured based on the number of publications produced by researchers individually or in groups, but other attributes can also be involved, such as where it is published, the number of authors and the number of citations, among others (Menéndez, Guerrero, Castellanos and Zurita, 2020). Commonly, this process is associated with one or more digital repositories: repositories of digital files that have different classifications and can be accessed, disseminated, and preserved (Texier, De Giusti, Oviedo, Villarreal, & Lira, 2012).

Normally, the analysis of researchers' publications is done manually. The results obtained from the search mechanism offered by the repository are considered, which keeps the metadata for each publication stored according to a standard description format (Chuttur, 2014). This entails the possibility of making errors in capturing information for the search, in the selection of scientific products or when interpreting the results (Cechinel, Sánchez and Sicilia, 2009). This margin of error can influence the characterization of the researchers' production.

In this sense, knowledge extraction techniques, specifically data mining, may be relevant to identify the behaviors of researchers associated with an institution. Thus, the objective of this research was to analyze the scientific production of the academics of an institution using data mining algorithms in order to identify the patterns and trends of their research activities.

As a case study, the scientific production of academics from the Autonomous University of Yucatan (UADY), an important public institution in southeastern Mexico, was used. In short, a search for new information was undertaken that could not normally be obtained in an analysis carried out manually.

State of knowledge

Scientific production and repositories

Something that characterizes contemporary science is the constant collaboration between scientists in multidisciplinary and transdisciplinary projects (González and Gómez, 2014). Without a doubt, this is an advantage to analyze concepts and conceive others. Needless to say, this group of people makes up what is known as the scientific community. Similarly, there are different groups of researchers in universities who share interests and carry out collaborative activities and which are known as academic bodies or research groups. These communities and bodies (as well as individual researchers) generate various types of scientific products: journal articles, books, presentations, among others, and can be analyzed from different perspectives (Guerrero, Menéndez, Castellanos and Curi, 2019).

This production is usually found in one or several digital repositories owned by the institution where the researcher is assigned (Guerrero, Menéndez and Castellanos, 2018). In Mexico there is the National Repository (https://www.repositorionacionalcti.mx/), which is defined as a digital platform in charge of providing open access to a wide variety of academic, scientific and technological information resources generated in Mexico. (National Council of Science and Technology [Conacyt], 2017). This repository integrates the institutional repositories (digital platforms of institutions belonging to the social, private and government sectors) and their respective authors.

Internationally, Scopus is one of the largest databases of citations and abstracts of peer-reviewed scientific literature: scientific journals, books, and conference proceedings (Elsevier, 2020). It has more than 70 million resources, 70,000 institution profiles and 16 million author profiles (Elsevier, 2019). It offers an exhaustive summary of the results of world research in various fields of science, which is why numerous institutions and organizations use it to know the productivity of their members through its indicators.

Scopus has achieved recognition due to the fact that it integrates in its journal indexes a significant number of titles corresponding to developing countries, combining both international, regional and local journals. (Luna, Luna y Luna, 2018).

The quality of the researcher in Mexico

The National System of Researchers (SNI) was created to recognize the various works done by people dedicated to the scientific and technological field in Mexico (Conacyt, 2019). The SNI presents three distinctions that a member can obtain (Conacyt, 27 de enero de 2017).

- *Candidate for National Researcher. He has products and publications in the scientific or technological fields.*
- *National Researcher. It is divided into three levels:*
    - Level 1. Has quality scientific or technological products, directs undergraduate or postgraduate theses, or teaches subjects, and participates in other teaching activities.
    - Level 2. In addition to what is necessary to belong to level I, collaborate with other researchers in original and quality products to demonstrate some line of research, as well as direct postgraduate theses and train human resources.
    - Level 3. In addition to what is necessary to belong to level II, it has research that causes an impact today, carries out activities of national leadership in science and technology, and has national and international recognition for its work.
- *National Researcher Emeritus.* To belong to this distinction, it is necessary that at the end of the call the candidate is 65 years old, at least 15 years assigned to the SNI and three uninterrupted evaluations obtaining the distinction National Researcher level 3.

Each applicant must comply with the guidelines established by the system, in addition to having a doctor's degree.

On the other hand, the Program for the Professional Development of Teachers for the Superior Type (Prodep) aims to professionalize full-time professors (PTC) in Mexico so that they form academic bodies and carry out teaching, research and development of technology and innovation activities. making use of social responsibility (General Directorate of Higher

University and Intercultural Education [DGESUI], 2014). The Prodep profile is awarded to those academics with a postgraduate degree who carry out research in addition to teaching and tutoring.

Prodep also considers the formation of research groups, called academic bodies, and classifies them into three groups (Professor Improvement Program [Promep], 2020). The characteristics presented below correspond to state and related universities, because the case study is applied with academic bodies of a university belonging to that group.

- Academic body in training (CAEF). They are academic bodies that are born from one or more lines of research and are at an early stage. Its characteristics are: 1) the members are identified, 2) at least half of its members have the Prodep profile, 3) they have defined the lines of generation or application of the knowledge that they will cultivate and 4) they have identified the academic bodies related to the who propose and high level to establish contact.
- Academic body in consolidation (CAEC). It is the intermediate level in which an academic body can be classified. It is characterized by: 1) more than half of its members have a doctorate, 2) they have quality academic products derived from consolidated lines of research, 3) at least a third of its members have the Prodep profile, 4) participate jointly in lines of research or application of knowledge, 5) extensive experience in teaching and training of human resources and 6) collaborate with other academic bodies.
- Consolidated academic body (CAC). It is the maximum level that an academic body can reach. Its characteristics are: 1) most of its members have a doctorate, 2) extensive teaching experience, 3) most of its members have a Prodep profile, 4) collaboration and scientific and academic production, 5) participate in congresses, seminars, tables, workshops, etc., on a regular and frequent basis and 6) intense participation in academic exchange networks.

In summary, the exposed Mexican instances, the SNI and Prodep, are options that Mexican researchers have to achieve a distinction for their work and that their research is used by the community in general. Each instance has its own indicators or guidelines to determine the quality of the academic and it is necessary that they be specific and objective so that anyone who aspires to a distinction knows if they meet the minimum necessary, based, in part, on the quality of their scientific investigations.

Data mining

Data mining is a field of statistics and computer science. Through various techniques, information is extracted from a database to generate knowledge, which can be expressed through concepts, rules, laws, patterns, among others (Romero and Ventura, 2020).

Data mining is a topic that involves practical learning, not so much theoretical (Witten, Frank and Hall, 2011); seeks techniques to find and describe structural patterns in data. It is a tool to help explain such data and make predictions from it. According to Romero and Ventura (2006), there are three basic techniques to discover patterns and knowledge:

1) Classification. It consists of determining new patterns based on a set of previously identified data. Some of the most used algorithms for this technique are: ID3, J48, C4.5, Naive Bayes, evolutionary algorithms, among others.

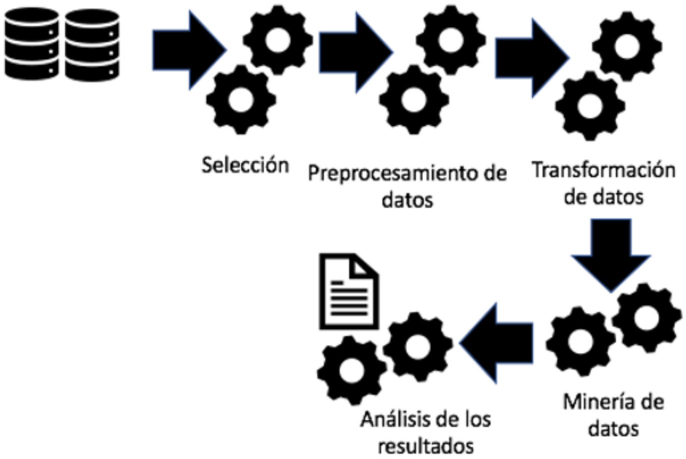
2) Grouping. Its main objective is to concentrate data that have similar characteristics. To do this, the data stored in the database is analyzed, and according to classification rules, a collection of resources grouped into classes is generated. Some of the most representative algorithms are: Single-link, Complete-link, SimpleKMeans, Kmedia, among others.

3) Association. Its main objective is to establish rules that associate the values of different attributes of the same database. Association and correlation are used to search for a frequent item among a large amount of information. Some of the most representative algorithms are: Apriori, Predictive A priori, among others.

Methodology

The methodology used was knowledge discovery in databases (KDD) (Guarascio, Manco and Ritacco, 2019). It consists of five phases (Camacho, Zapata, Menéndez and Canto 2018), which are described below (figure 1).

Figura 1. Fases de la propuesta metodológica



Fuente: Elaboración propia

1) *Selection.* It is made up of two sub-phases. The first consists of learning the domain of knowledge, especially that which is relevant and the goals of the application. The second is to select the target databases. In this case, those that store scientific production and those that contain information from researchers.
2) *Preprocessing.* It consists of using basic operations that allow purging the data that is not required, selecting the necessary and those that could be useful. In this case, the elementary data of each researcher and the descriptive data of each publication (title, authors, keywords, identifiers, among others).
3) *Transformation.* In this stage, various fields of the numerical type database are transformed into linguistic labels for a better characterization. Some useful techniques are the use of percentiles to associate ranges with keys.
4) *Data mining.* Various data mining techniques are used, according to the needs of the problem posed. For each technique, related algorithms are studied and the appropriate one is selected in each case. Among the techniques used are: classification, grouping and association rules.
5) *Analysis of the results.* The generated results are examined, which concludes in the generation of new knowledge regarding academic bodies and researchers. From this, it is possible to carry out decision making.

It is important to point out that data collection instruments are not used, given the origin of the information that will be analyzed. The data is collected automatically through a web application that accesses the information stored in the repositories, and is then processed according to the methodology described.

Case study

UADY is one of the most important higher education institutions in southeastern Mexico. Until February 2019, 824 full-time professors and 78 academic bodies were registered at UADY, distributed in 15 faculties and two research centers that are grouped into six campuses (UADY, 2020). Table 1 presents some statistics that UADY professors have reported as of February 2019. For the purpose of this work, we will call a full-time professor who has production a researcher.

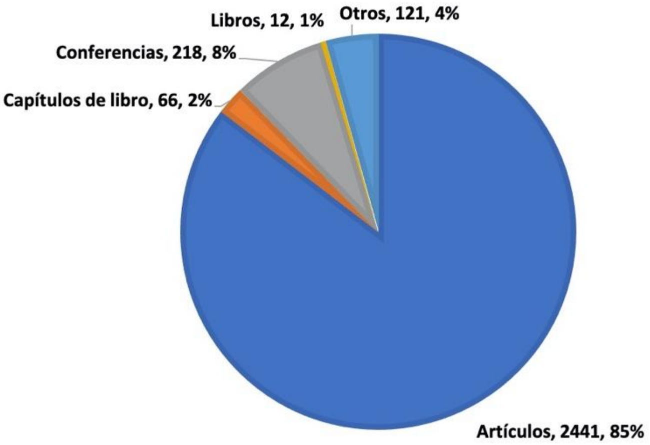
Tabla 1. Algunas estadísticas de los profesores de la UADY

Profesores con producción	438
Doctores	280
Profesores en el SIN	193
Doctores con SIN	64.5 %
Producción promedio de los doctores	15.5 publicaciones
Producción promedio de doctores SNI	19.73 publicaciones
Doctores en cuerpos académico7.85 %	
Doctores SNI pertenecientes a un cuerpo académico	78.75 %

Fuente: Elaboración propia

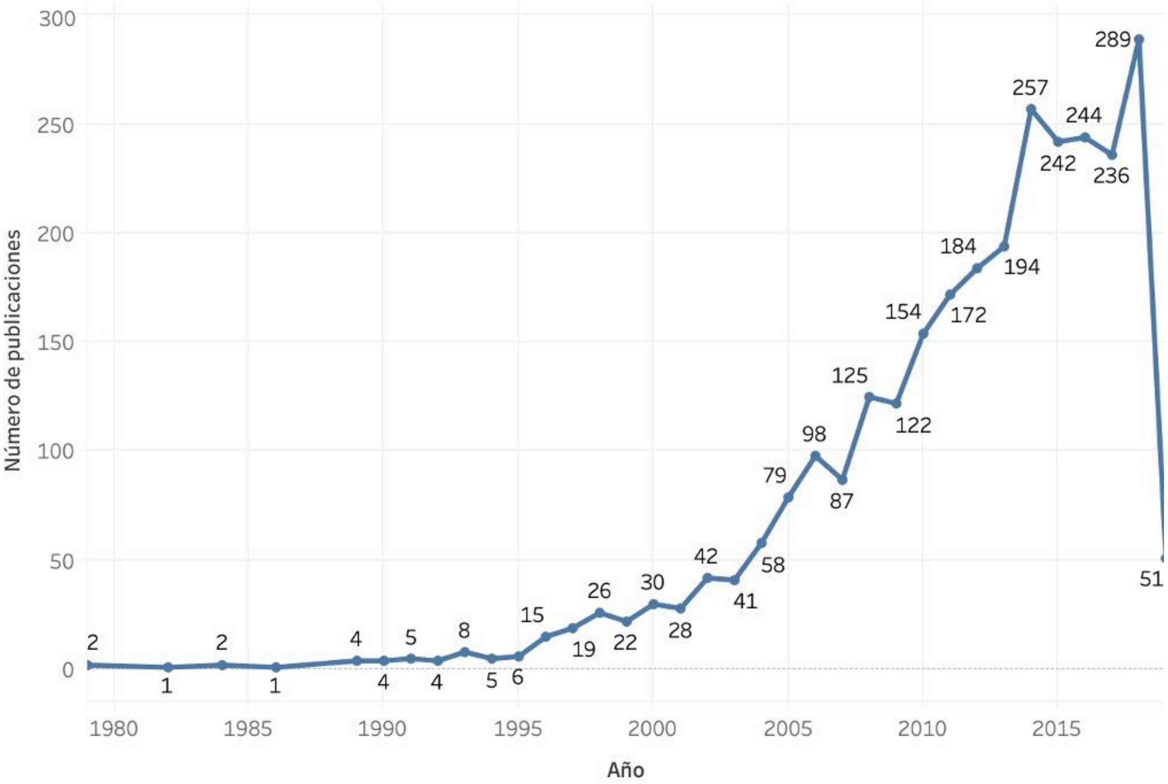
Figures 2 and 3 show the distribution of the UADY production stored in Scopus and the national repository in the period from the first publication that turned out to be in Scopus (1979) to February 2019.

Figura 2. Distribución de la producción científica de la UADY por tipo



Fuente: Elaboración propia

Figura 3. Producción científica de la UADY por año



Fuente: Elaboración propia

Selection

The information was obtained through its own web application, based on the Python programming language (McKinney, 2017), which retrieves the scientific production of the UADY through a Scopus query interface and the National Repository. It uses a database in MongoDB (MongoDB, 2019) to also store the relevant information of the 78 academic bodies. Table 2 describes the attributes used for academic bodies.

Tabla 2. Atributos para cuerpos académicos

Atributo	Descripción
Nombre	Nombre del cuerpo académico
número_de_integrantes	Número de integrantes
Facultad	Facultad a la que pertenece
Campus	Campus al que pertenece
Tipo	Tipo de cuerpo académico: CAEF: cuerpo académico en formación CAEC: cuerpo académico en consolidación CAC: cuerpo académico consolidado
artículos_indizados	Número de artículos indizados por Scopus
capítulos_indizados	Número de capítulos indizados por Scopus
libros_indizados	Número de libros indizados por Scopus
otros_indizados	Número de productos de otra índole indizados por Scopus
producción_repositorio	Número de productos almacenados en el Repositorio Nacional
total_de_publicaciones	El total de producción indizada y no indizada de los cuerpos académicos

Fuente: Elaboración propia

Table 3 presents the distribution of academic bodies by campus and the average production indexed by Scopus and the National Repository.

Tabla 3. Distribución de los cuerpos académicos en los campus de la UADY y la producción promedio

Campus	CAC	Producción promedio	CAEC	Producción promedio	CAEF	Producción promedio
Campus de Arquitectura, Hábitat y Diseño	1	0	2	0	0	N/A
Campus de Ciencias Biológicas y Agropecuarias	8	10.3	2	12.5	0	N/A
Campus de Ciencias de la Salud	4	8.75	8	3.62	1	0
Campus de Ciencias Exactas e Ingenierías	10	11.88	9	7.22	2	1
Campus de Ciencias Sociales, Económico Administrativas y Humanidades	9	3.44	9	0	3	0
Centro de Investigaciones Regionales Dr. Hideyo Noguchi	6	19.66	2	17	2	1.5

Fuente: Elaboración propia

The attributes listed in Table 4 were retrieved for the 438 professors who have scientific production.

Tabla 4. Atributos para profesores

Atributo	Descripción
Prodep	Verifica si el profesor cuenta con perfil Prodep (sí o no)
último_grado	Último grado de estudios (licenciatura, maestría o doctorado)
Género	Género (masculino o femenino)
Sin	Nivel de SNI con el que cuenta el profesor (candidato, nivel 1, nivel 2, nivel 3 o no)
cuerpo_académico	Cuerpo académico al que pertenece
Facultad	Facultad a la que pertenece
Campus	Campus al que pertenece
Cuartil	Cuartil de citas al que pertenece
Activo	Verifica si el profesor se encuentra activo. Para ello, deberá tener al menos tres publicaciones en los tres años anteriores.
total_scopus	Número total de publicaciones indizadas por Scopus
total_repositorio	Número total de publicaciones indizadas por el Repositorio Nacional

Fuente: Elaboración propia

Table 5 shows the groups of researchers according to their academic degree, as well as their distribution on campus.

Tabla 5. Distribución de los profesores con producción en los campus de la UADY

	Campus de Ciencias Biológicas y Agropecuarias	Campus de Ciencias de la Salud	Campus de Ciencias Exactas e Ingenierías	Campus de Ciencias Sociales, Económico Administrativas y Humanidades	Centro de Investigaciones Regionales Dr. Hideyo Noguchi
Profesores con producción	76	86	159	54	63
Doctores	59	34	98	41	47
Profesores con perfil Prodep	75	84	157	54	61
Profesores no adscritos al SIN	39	68	94	17	27
Candidato al SIN	0	6	10	10	2
SNI 1	25	9	46	19	25
SNI 2	6	3	7	7	9
SNI 3	6	0	2	1	0
Profesores SNI con cuerpo académico	33	13	54	23	29

Fuente: Elaboración propia

Preprocess

To carry out the data mining process, the developed system generates .csv and .arff files from the database. The files compile the relevant information for the purpose of the investigation: for the academic bodies, list the name, number of members, campus, type, production indexed by the two repositories and total number of publications; for professors, the last grade, Prodep profile, gender, SNI level, academic body, campus, total publications, production indexed by the two repositories, citation quartile and activity of each one are listed.

Transformation

Some numerical attributes for academic bodies and professors were categorized. This in order to classify the data for use in the next phase of the process. The main ones are described below.

For academic bodies:

- *Campuses.* This is a new attribute. To obtain it, the "faculty" attribute was used, where, depending on the faculty of the academic body, the corresponding campus was assigned.
- *Total publications.* Categories were defined depending on the number of total publications of an academic body. The labels are:
    - Very little: from 0 to 20 publications.
    - Few: from 21 to 41 publications.
    - Regular: from 42 to 62 publications.
    - A lot: from 63 to 83 publications.
    - Too much: more than 84 publications.

For teachers:

- *Quartile.* Assign a letter of the alphabet depending on the number of citations a scholar has:
    - From zero to two appointments.
    - Three to five appointments.
    - From six to nine appointments.
    - From 10 to 19 appointments.
    - From 20 to 29 appointments.

  - From 30 to 39 appointments.
  - From 40 to 79 appointments.
  - From 80 to 932 citations.
  - More than 933 citations.
- *Total Scopus.* There is a categorization in order to assign a value depending on the number of publications indexed by Scopus. These values are:
  - Very little: from 0 to 30 publications.
  - Few: from 31 to 60 publications.
  - Regular: from 61 to 90 publications.
  - A lot: from 91 to 120 publications.
  - Too much: more than 121 publications.
- *Total repository.* In order to have a better data representation, a scale was used and thus a label could be assigned depending on the number of publications of a researcher indexed by the National Repository. These labels are:
  - Very little: from zero to one publication.
  - Little: two to three publications.
  - Regular: four to five publications.
  - A lot: six to seven publications.
  - Too much: eight to nine publications.

It is important to mention that the average total production of the academic bodies analyzed is in the range of 0 to 20 (label “Very little”), and the same occurs with researchers (0 to 30, label “Very little”).

Data mining

For the analysis of the files, the WEKA software (Hall et al., 2009) was used. In this tool, the J48 (classification), SimpleKMeans (grouping) and A priori (association) algorithms (Witten et al., 2011) were applied and a collection of classification trees, patterns and rules was obtained, which will be explained later.

Analysis of the results of data mining techniques

In this phase, the results generated with the three previously presented data mining techniques were analyzed.

Results

This section contains the results of the implementation of the data mining (with the help of the classification, grouping and association algorithms) with the WEKA software from the information generated from the production of academic bodies and researchers. In the first instance, the algorithms were applied to the academic bodies of the UADY, and in the second instance, the same was done for its academics. The order followed was to apply classification algorithms to generate a classification tree, then clustering algorithms to divide both academic and research bodies into groups, and finally extract rules with the help of the association algorithm. Below are the results obtained.

Classification algorithm

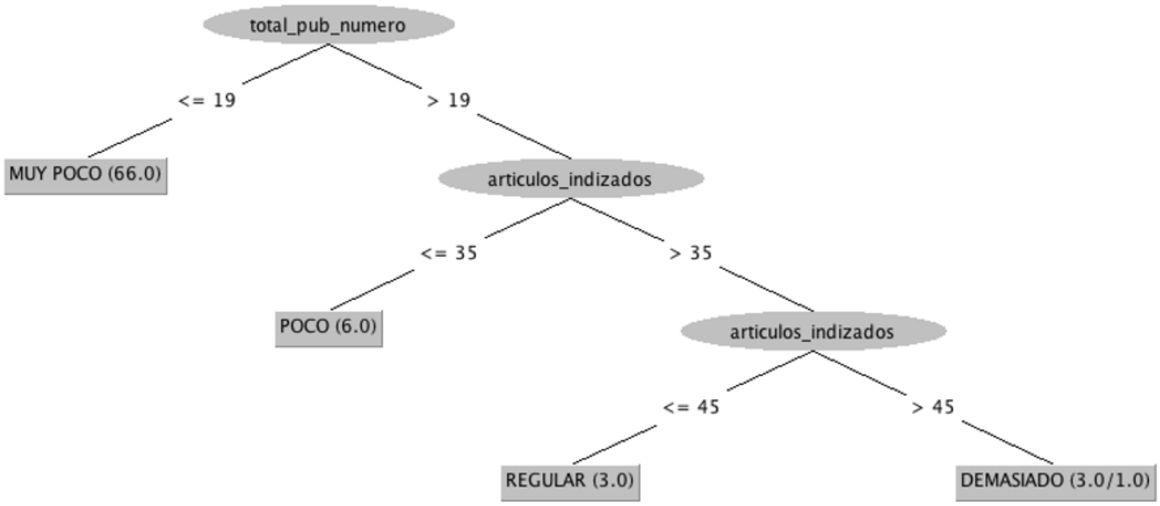
The problem of classifying individuals or entities has been of great interest for research (Romero and Ventura, 2010, 2020). In this work, the J48 algorithm was experimented with.

First, it was validated with the academic bodies of the university. To do this, the total number of publications was used as the main attribute, which ranged from 0 to 103 publications. The J48 algorithm for this experiment has a degree of correctness of 91.0256%. Once applied, the following results were obtained (represented in Figure 4):

- If the total number of publications is less than or equal to 19, then they are very few.
- If the total number of publications is greater than 19 and less than or equal to 35, being indexed articles, then there are few.
- If the total number of publications is greater than 35 and less than or equal to 45, being indexed articles, then it is regular.
- If the total number of publications is greater than 45, being indexed articles, then there are too many.

Analyzing these results, it can be seen that the algorithm has discarded the classification "A lot" and "Too much", as well as other resources indexed by Scopus or the National Repository, due to the minimum amount they represent with respect to the total production. analyzed (figure 4).

Figura 4. Árbol de clasificación J48 para el total de publicaciones de los cuerpos académicos de la UADY



Fuente: Elaboración propia

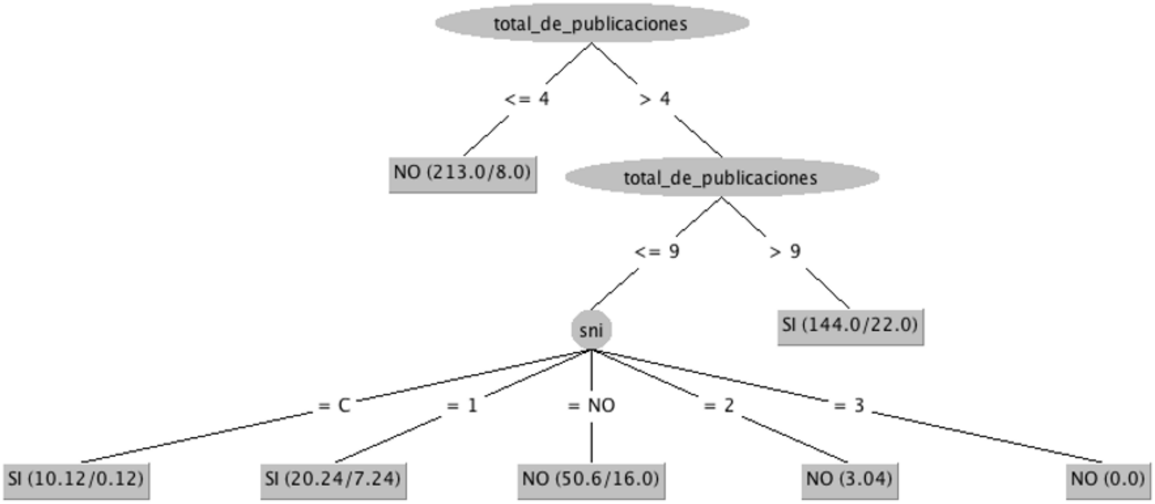
In the second instance, researchers from the university were experimented with. For this, the main attribute was used if the teacher is active or not and the production generated in the period from January 2016 to January 2019. For the purposes of the experiment, a teacher is considered to be active if he has at least three publications from January 2016 to January 2019.

The J48 algorithm for this experiment has a degree of correctness of 84.3537%. When applied, the following results were obtained (represented in figure 5):

- If a teacher's total posts is less than or equal to four, then they are not active.
- If a professor's total publications is greater than four and less than or equal to nine, and he is also a candidate for SNI, then he is active.
- If a professor's total publications is greater than four and less than or equal to nine, and he also has SNI level 1, then he is active.
- If a professor's total publications is greater than four and less than or equal to nine, and he also has SNI level 2, then he is not active.
- If a professor's total publications is greater than four and less than or equal to nine, and he also has SNI level 3, then he is not active.
- If the total number of publications of a teacher is greater than four and less than or equal to nine, and also does not have SNI, then it is not active.
- If a teacher's total posts is greater than nine, then they are active.

For SNI levels 2 and 3, there are not enough cases, so the J48 algorithm considers teachers with these levels as not active (figure 5).

Figura 5. Árbol de clasificación J48 para el total de publicaciones de profesores de la UADY

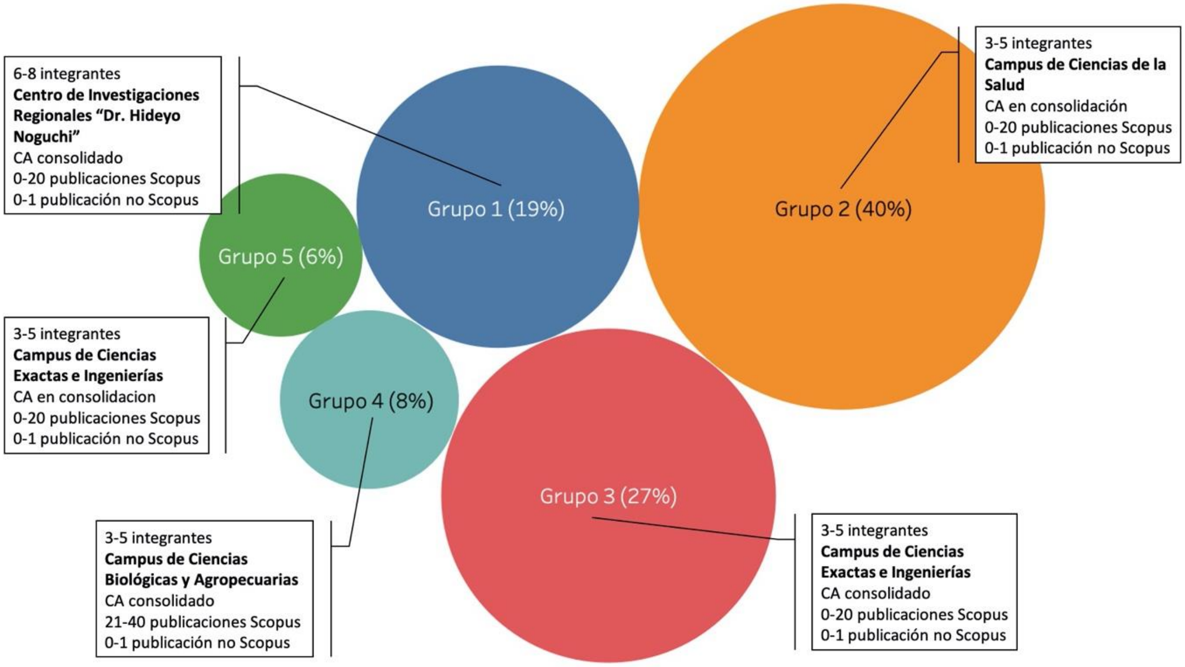


Fuente: Elaboración propia

Clustering algorithm

SimpleKMeans was experimented with as a clustering algorithm. For this, it was tried to create five groups of academic bodies, hoping that the separation of the groups would be based on the campuses that the university has. However, this was not the case (figure 6), since several academic bodies were not considered relevant to the algorithm and were absorbed by others with greater weight.

Figura 6. Generación de cinco grupos de cuerpos académicos con el algoritmo SimpleKMeans



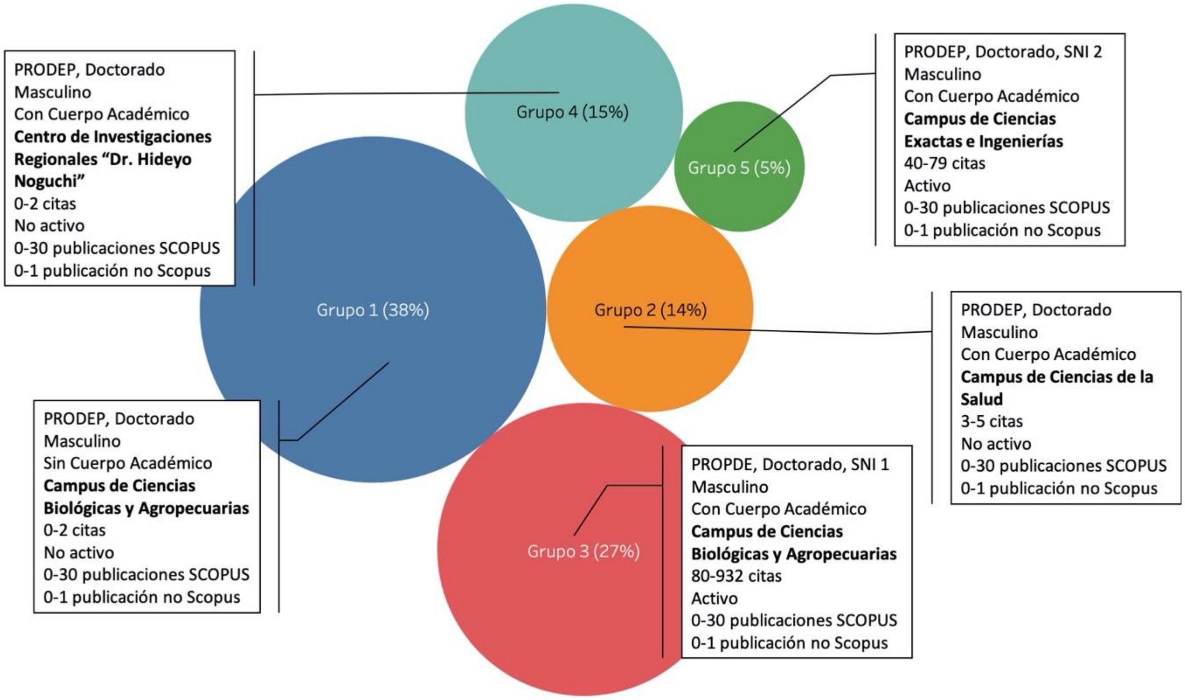
Fuente: Elaboración propia

The representative groups reported in figure 6 indicate a high presence of consolidated academic bodies (54%), followed by academic bodies in consolidation (46%). The number of members is mostly small (between three and five [81%]) with the exception of group 1 (which has between six and eight members). Although almost all the groups have a production between 0 and 20 publications indexed in SCOPUS, group 4 (8%) has a range of 21-40 publications in Scopus, they are consolidated and belong to the Campus of Biological and Agricultural Sciences. The most significant group of academic bodies (32 [40%]) is in the process of consolidation with three to five members, belongs to the Health Sciences Campus and has between 0 and 20 Scopus publications.

In the second instance, it was experimented in the same way with the creation of five representative groups of researchers hoping that they would be grouped by campus. Again, this was not the case, since the academics of the Campus of Biological and Agricultural Sciences have a greater production, which is why they prevail over academics from other campuses and cause the latter to not be relevant for the algorithm.

Figure 7 shows that the most significant group (38%) groups together doctors who have a Prodep profile without an academic body, whose citation level is very low and a number between 0 and 30 publications. In two groups, the professors publish each year, since they have SNI, between 0 and 30 publications, with a high level of citations (80 to 932 citations for group 3 [27%]; 40 to 79 citations for group 5 [5%].]). All the representative groups of researchers have Prodep and a doctorate as their last degree of studies, but most do not have SNI (three groups).

Figura 7. Generación de cinco grupos de profesores con el algoritmo SimpleKMeans



Fuente: Elaboración propia

Association algorithms

The generation of association rules is the last data mining technique used in this case study. The Apriori algorithm was used to observe the behavior of the academic bodies and professors of the UADY. For this, 25 rules for professors and 25 rules for academic bodies were generated, of which the 10 most reliable are presented (table 6).

Rules have one or more antecedents that generate a consequent. For each component of a rule, the number of cases that were considered to generate it is established, which gives its confidence index.

Many of the rules confirm assumptions or the existing correlation between attributes. For example, the number of members of an academic body, the level of certification achieved and their productivity or among researchers, their degree, the Prodep profile and their activity.

Tabla 6. Reglas generadas con el algoritmo Apriori

Antecedente	Consecuente	Interpretación en lenguaje natural	Índice de confianza

no_integrantes = Pocos
tipo = CAEC
(28)
	total_publicaciones = Muy poco
	(28)
		Si el cuerpo académico está en consolidación y el número de integrantes es poco, entonces el grupo tiene muy poca producción.	1

último_grado = D (279)	prodep = Sí (278)	Si el último grado del investigador es doctorado, entonces cuenta con perfil Prodep.	1
activo = No (272)	total_scopus = Muy poco (271)	Si el investigador no se encuentra activo, entonces tiene muy pocas publicaciones en Scopus.	1

total_socups = Muy poco
total_repositorio = Muy poco
(356)
	prodep = Sí
	(349)
		Si un investigador cuenta con muy pocas publicaciones Scopus y Repositorio Nacional, entonces cuenta con perfil Prodep.	0.98

tipo = CAEC (32)	total_publicaciones = Muy poco (31)	Si el cuerpo académico está en consolidación, entonces tiene muy poca producción.	0.97

producción_scopus = Muy poco (66)	producción_repositorio = Muy poco (62)	Si la producción en Scopus de un cuerpo académico es muy poca, entonces su producción en el Repositorio Nacional será muy poca.	0.94
no_integrantes = Pocos (61)	producción_scopus = Muy poco (56)	Si el número de integrantes del cuerpo académico es poco, entonces tiene muy poca producción en el Repositorio Nacional.	0.92

no_integrantes = Pocos
produccion_repositorio = Muy poco
(58)
	producción_scopus = Muy poco
	(53)
		Si el número de integrantes del cuerpo académico es poco y tiene muy poca producción en el Repositorio Nacional, entonces tiene muy poca producción en Scopus.	0.91

prodep = Sí
total_repositorio = Muy poco
(382)
	total_scopus = Muy poco
	(349)
		Si un investigador cuenta con perfil Prodep y cuenta con muy pocas publicaciones en el Repositorio Nacional, entonces cuenta con muy pocas publicaciones en Scopus.	0.91

prodep = Sí (434)	total_scopus = Muy poco (395)	Si el investigador cuenta con perfil Prodep, entonces cuenta con muy pocas publicaciones en Scopus.	0.91

Fuente: Elaboración propia

Discussion

For the classification algorithm, it was possible to observe a cut in the total publication of the academic bodies. It only focused on an interval from 0 to 45, with which the different labels related to the number of publications of an academic body were obtained. This fact is reinforced by the results obtained through an index system (Guerrero, Menéndez and Castellanos, 2021) for the evaluation of the production of an academic body and a researcher, since the average of the publications of an academic body is of 11.25, and most of these groups have up to 19 posts.

While, for teachers, it was observed that the J48 tree is much more specific in the results obtained, which suggests that authors with publications less than or equal to four tend to be inactive. This is probably because the production it has was made prior to the previous three years. While if they are greater than nine they tend to be active; In the interval between these two values, tendencies to not have SNI can be observed, which, as previously mentioned, is due to the small amount of data from level 2 and 3 teachers.

The case of active researchers is complemented by the results obtained through an ontological model for the representation of knowledge in the domain of scientific production (Guerrero, Menéndez, Castellanos and Gómez, 2019). Through a SPARQL query engine, said study identified that the majority of active professors belong to the SNI at level 1. Therefore, with what is obtained in the classification tree, it is inferred that these professors have more than nine publications in the last three years.

It is worth mentioning that the decision tree of the academic bodies considers all the academic bodies, even those whose production could not be located. Instead, the teachers' decision tree excludes all those teachers who do not have production.

For the grouping algorithm, it was possible to observe the great impact of the Health Sciences, Biological and Agricultural Sciences and Exact Sciences and Engineering campuses, since they represent a very large percentage of the results obtained once classified into groups. These results were presented even using a greater number of groups or clusters, which implies that the other campuses do not have a large contribution of scientific articles for the university.

Even in another work where collaborations between UADY professors were studied (regardless of whether they are academic bodies or not) (Guerrero, Menéndez, Castellanos and Curi, 2020), the existence of collaborations in the university dependencies was confirmed, even between them, generating multidisciplinary knowledge. The previously mentioned ontological model inferred that there are publications that cover up to six areas of knowledge from seven different ones.

Similarly, a tendency can be seen to have consolidated academic groups and professors with a doctorate degree, this reinforces Prodep's requirements for consolidated academic bodies, where it is requested that the majority of the members of an academic body must have doctorate. They even corroborate the results of another study on the production and collaboration of academic bodies (Guerrero, Menéndez, Castellanos and Guerra 2020), which indicates that, the greater the consolidation, the greater the number of collaborating national and international institutions. This in turn confirms the degree of consolidation that has been assigned to them, since the definition of ties with external institutions to generate and disseminate new knowledge is part of the evaluation criteria.

Although, in the same way, it can be seen that, both in academic bodies and in researchers, there is a tendency to have very little production, both indexed and non-indexed. It is considered very little because the maximum value in this variable is very high. One of the most interesting data is the tendency in the groups of professors to have almost mostly male professors, which gives an idea of the difference that exists in the professors within the university.

Finally, for the association algorithm, the rules generated for the academic bodies reflect trends proportional to the number of members of a body with its production. While different rules were obtained for professors, such as the relationship between the Prodep profile with the last degree of studies, gender and number of publications, as well as proportional rules with the number of indexed and non-indexed publications, this reinforces compliance. in the requirements of the Prodep profile and the SNI.

The tendency of researchers to have very few publications is reinforced by the use of an index system for the evaluation of the production of academic bodies and researchers (Guerrero, Menéndez and Castellanos, 2021), since, in the results obtained , it is indicated that in the UADY the average number of publications is around 11 (and very few researchers have more than 121 publications in Scopus, which is considered a very high number of publications in said bibliographic database).

Conclusions

The digital repositories of scientific documentation have made information about the publications made by researchers available to anyone, helping to disseminate and take advantage of scientific and technological advances. Each resource stored in a digital repository can be described, located and referenced through its metadata. Its analysis generates relevant information for decision making.

Various instances such as the SNI and Prodep, both in Mexico, are responsible for evaluating and granting recognition to researchers through their scientific production stored in various repositories.

This paper has presented how data mining techniques allow researchers to be characterized at the group or individual level of an institution. For this, a group of indicators was defined that considers the number of publications, citations, the prestige of the journals and, to a lesser extent, non-indexed and open access production.

The results obtained would allow corroborating or discarding assumptions related to the scientific productivity of an institution, which facilitates decision-making that encourages or conditions institutional policies.

Future lines of research

For future research, it is intended to corroborate the results obtained through the data mining methodology with new data in particular situations, such as, for example, characterizing the professors and academic bodies at the campus level or areas of knowledge, which would allow the identification of professors and groups with similar production profiles for possible collaborations.

It is intended to carry out a new study that incorporates the impact of the publication of an academic body and a researcher for each citation received. Each citation would be valued according to the prestige of the original journal or publication, including self-citations. This would allow the characterization of the dissemination of the generated knowledge.

The work proposal allows building the basis for new complementary studies where bibliometrics, together with computer science, allows evaluating the impact of scientific publications in society in general, and not only in the scientific community, since the Resources found in multiple repositories are referenced on other platforms with a social and academic focus.

In addition, in the case of universities and research centers that offer postgraduate programs, it is expected to characterize, by line of research, the most frequent topics and, through statistical indicators, identify the most relevant ones; thus facilitating various activities related to the dissemination and promotion of said programs.

On the other hand, it is intended to extend the panorama from one to several institutions, which will allow the characterization of collaborations between academic bodies by geographical and thematic areas.

It is important to highlight the values of the proposed indicators and the importance of their calculation through an automatic process capable of analyzing the current state of an institution's research, covering aspects that can hardly be obtained in a manual process. In this sense, we are working on software that implements the methodology and algorithms presented. The application will use an architecture based on web services and protocols for interoperability, which will facilitate its extensibility so that it can be easily adapted according to new needs.

References

Camacho, P. E., Zapata, A., Menéndez, V. H. y Canto, P. J. (2018). Análisis del desempeño del profesorado universitario en el uso de MOODLE a través de técnicas de minería de datos: propuestas de necesidades formativas. *RED. Revista de Educación a Distancia*, (58). Recuperado de https://doi.org/10.6018/red/58/10.

Cechinel, C., Sánchez, S. y Sicilia, M. Á. (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. Paper presented at the Research Conference on Metadata and Semantic Research. Milan, October 1-2, 2009. Retrieved from https://doi.org/10.1007/978-3-642-04590-5_6.

Chuttur, M. Y. (2014). Investigating the effect of definitions and best practice guidelines on errors in Dublin Core metadata records. *Journal of Information Science*, *40*(1), 28-37. Retrieved from https://journals.sagepub.com/doi/10.1177/0165551513507405.

Consejo Nacional de Ciencia y Tecnología [Conacyt]. (27 de enero de 2017). Acuerdo por el que se emite el nuevo reglamento del Sistema Nacional de Investigadores. *Diario Oficial de la Federación.* Recuperado de http://www.dof.gob.mx/nota_detalle.php?codigo=5470107&fecha=27/01/2017.

Consejo Nacional de Ciencia y Tecnología [Conacyt]. (2017). Lineamientos jurídicos de ciencia abierta. Recuperado de https://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-juridicos-de-ciencia-abierta/3828-lineamientos-juridicos-de-ciencia-abierta/file.

Consejo Nacional de Ciencia y Tecnología [Conacyt]. (2019). Sistema Nacional de Investigadores. Recuperado de https://www.conacyt.gob.mx/index.php/el-conacyt/sistema-nacional-de-investigadores.

Dirección General de Educación Superior Universitaria e Intercultural [DGESUI]. (2014). Programa para el Desarrollo Profesional Docente, para el Tipo Superior (Prodep). Recuperado de https://www.dgesui.ses.sep.gob.mx/PRODEP.htm.

Elsevier. (2019). Scopus. Expertly curated abstract & citation database Retrieved from https://www.elsevier.com/solutions/scopus.

Elsevier. (2020). *Scopus*. *Content Coverage Guide.* Elsevier. Retrieved from https://www.elsevier.com/solutions/scopus/how-scopus-works/content.

González, G. y Gómez , J. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista Española de Documentación Científica*, *37*(4). Recuperado de https://doi.org/10.3989/redc.2014.4.1186.

Guarascio, M., Manco, G. y Ritacco, E. (2019). Knowledge Discovery in Databases. In Ranganathan, S., Gribskov, M., Nakai, K. and Schönbach, C. (eds.), *Encyclopedia of Bioinformatics and Computational Biology* (vol. 1) (pp. 336-341). United States: Elsevier. Retrieved from https://doi.org/10.1016/B978-0-12-809633-8.20456-1.

Guerrero, J. D. T., Menéndez, V. H. y Castellanos, M. E. (2018). Indicadores de calidad en investigaciones científicas: antecedentes. *Abstraction and Application*, *19*(9), 6-24. Recuperado de https://intranet.matematicas.uady.mx/journal/descargar.php?id=134.

Guerrero, J. D. T., Menéndez, V. H. y Castellanos, M. E. (2021). An indexing system for the relevance of academic production and research from digital repositories and metadata. *The Electronic Library*, *39*(1), 33-58. Retrieved from https://doi.org/10.1108/EL-06-2020-0160.

Guerrero, J. D. T., Menéndez, V., Castellanos, M. E. y Curi, L. F. (2019). Use of Graph Theory for the Representation of Scientific Collaboration. Paper presented at the International Conference on Computational Collective Intelligence, Hendaye, September 4-6, 2019. Retrieved from https://doi.org/10.1007/978-3-030-28374-2_47.

Guerrero, J. D. T., Menéndez, V. H., Castellanos, M. E. y Curi, L. F. (2020). Analysis of Internal and External Academic Collaboration in an Institution Through Graph Theory. *Vietnam Journal of Computer Science*, *7*(4), 391-415. Retrieved from https://doi.org/10.1142/S2196888820500220.

Guerrero, J. D. T., Menéndez, V. H., Castellanos, M. E. y Gómez, J. R. (2019). Use of an Ontological Model to Assess the Relevance of Scientific Production. *IEEE Latin America Transactions*, 17(9), 1424-1431. Retrieved from https://doi.org/10.1109/TLA.2019.8931135.

Guerrero, J. D. T., Menéndez, V. H., Castellanos, M. E. y Guerra, C. A. (2020). Metodología para la generación de grafos de colaboración científica de una institución académica. *Abstraction and Application*, *29*, 88-101. Recuperado de https://intranet.matematicas.uady.mx/journal/descargar.php?id=208.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18. Retrieved from https://doi.org/10.1145/1656274.1656278.

Leahey, E. (2016). From Sole Investigator to Team Scientist: Trends in the Practice and Study of Research Collaboration. *Annual Review of Sociology*, *42*(1), 81-100. Retrieved from https://doi.org/10.1146/annurev-soc-081715-074219.

Luna, M., Luna, E. y Luna, S. (2018). La UADY en la literatura científica registrada en Web of Science y Scopus: 1900-2016. *Revista Educación y Ciencia*, *7*(50), 17-29. Recuperado de http://www.educacionyciencia.org/index.php/educacionyciencia/article/view/470.

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. United States: O'Reilly Media.

Menéndez, V. H., Guerrero, J. D. T., Castellanos, M. E. y Zurita, E. (2020). Análisis de la producción de cuerpos académicos basado en teoría de grafos. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *10*(20). Recuperado de https://doi.org/10.23913/ride.v10i20.603.

MongoDB. (2019). The Database for Modern Applications. Retrieved from https://www.mongodb.com/.

Programa de Mejoramiento del Profesorado [Promep]. (2020). Cuerpos académicos. Conceptos básicos. Recuperado de http://promep.sep.gob.mx/ca1/Conceptos2.html.

Romero, C. y Ventura, S. (eds.) (2006). *Data Mining in E-Learning*. Southampton, United Kingdom: WIT Press.

Romero, C. y Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618. Retrieved from https://doi.org/10.1109/TSMCC.2010.2053532.

Romero, C. y Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3). Retrieved from https://doi.org/10.1002/widm.1355.

Texier, J., De Giusti, M. R., Oviedo, N., Villarreal, G. L. y Lira, A. J. (2012). El uso de repositorios y su importancia para la educación en Ingeniería. Ponencia presentada en el World Engineering Education Forum (WEEF). Buenos Aires, 2012. Recuperado de http://sedici.unlp.edu.ar/handle/10915/22943.

Universidad Autónoma de Yucatán [UADY]. (2020). Nuestra universidad. Quiénes somos. Recuperado de https://www.uady.mx/nuestra-universidad/.

Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington, United States: Elsevier. Retrieved from https://doi.org/10.1016/C2009-0-19715-5.

Rol de Contribución	Autor (es)
Conceptualización	Víctor Hugo Menéndez Domínguez
Metodología	Jared David Tadeo Sosa
Software	José William Cervera Pérez
Validación	María Enriqueta Castellanos Bolaños
Análisis Formal	María Enriqueta Castellanos Bolaños
Investigación	Víctor Hugo Menéndez Domínguez
Recursos	Jared David Tadeo Sosa
Curación de datos	José William Cervera Pérez
Escritura - Preparación del borrador original	Jared David Tadeo Sosa
Escritura - Revisión y edición	María Enriqueta Castellanos Bolaños
Visualización	María Enriqueta Castellanos Bolaños
Supervisión	Víctor Hugo Menéndez Domínguez
Administración de Proyectos	Víctor Hugo Menéndez Domínguez
Adquisición de fondos	Víctor Hugo Menéndez Domínguez