Capítulo 7

Minería de datos aplicada al análisis bibliométrico. Descripción y usos de reglas de asociación y modelos de regresión basados en árboles

José Hernando Ávila-Toscano

Introducción

El estudio del comportamiento de la ciencia se vale de un amplio número de métodos y técnicas mediante los cuales se busca ofrecer resultados fiables y lo suficientemente robustos para explicar patrones completos de funcionamiento, estructuras de organización científica y determinantes relacionados con la producción de conocimiento, entre otros intereses del campo cienciométrico.

Con el crecimiento exponencial de los productos de nuevo conocimiento publicados en formato digital, y el creciente número de revistas dentro de sistemas de indexación internacional y regional, resulta prácticamente imposible pensar en la realización de análisis de la información de estos productos desde las tradicionales metodologías bibliotecológicas y

estadísticas. De allí que cada vez sea más útil y necesario apelar a recursos computacionales y de inteligencia artificial para obtener el mejor alcance dentro del estudio de los productos del conocimiento, las dinámicas participativas en la ciencia, los impactos de la investigación y la constitución misma de campos de estudios a nivel disciplinar y subdisciplinar.

Hoy por hoy, la cienciometría se nutre de la riqueza metodológica del Análisis de Redes Sociales para la identificación de estructuras de cooperación emergentes entre investigadores e instituciones (véase capítulo 5); la aplicación de Modelos autologísticos de atributos de actor (ALAAM, por su sigla en inglés) para identificar los efectos de variables exógenas sobre los atributos de los actores de una red (Letina, 2016); el uso del método de co-términos para la detección de vinculaciones entre palabras que permite construir asociaciones semánticas a fin de vislumbrar campos temáticos en una disciplina científica (véase capítulo 6); el aporte del *Eigenfactor* como métrica susceptible de identificar (y diferenciar con otros índices) la popularidad y el prestigio de las publicaciones científicas (Franceschet, 2010); el desarrollo del método baricéntrico (*Barycenter method*) como medida de internacionalización de los productos científicos utilizando información geográfica o espacial (Verleysena & Engels, 2014); y tantos otros ejemplos citables que denotan la riqueza analítica en cienciometría y la relevancia del uso de algoritmos y sistemas de cómputo que aprovechan la Web para la recogida de volúmenes amplios de información.

Este capítulo concentra su interés en el método de Minería de datos (*Data mining*) como recurso de análisis cuantitativo que se surte de procedimientos de extracción de información, con amplias aplicabilidades en el campo de la cienciometría. Dos procedimientos de minería son privilegiados en este documento, las reglas de asociación y los árboles de

clasificación y regresión, por su simpleza, agilidad, eficacia y utilidad en la identificación de determinantes relacionados con la producción científica.

Inicialmente realizaremos una descripción concisa de lo que se entiende por Minería de Datos, y posteriormente se describirán los procedimientos de interés; para ganar contextualización de estos recursos, nos valdremos de la presentación de algunos resultados de investigación "extraídos" por medio de ejercicios de minería, aspirando con ello, ilustrar la utilidad y ventajas de estos procedimientos en los estudios de la producción científica.

7.1 Bases de datos y descubrimiento de conocimiento: el método de Minería

Como mencionábamos en la Introducción, el nivel actual de generación de productos de conocimiento es humanamente inconmensurable. Con el desarrollo de la Web 2.0 y la tendencia a la digitalización de la ciencia, la mayor parte de la información científica se encuentra en Internet, reposando en bases de datos que ofrecen acceso desde todos los rincones del planeta. Hoy en día, todos los investigadores del mundo tienen la oportunidad de interactuar en tiempo real con actores de los sistemas científicos y tecnológicos de todos lados, lo que amplía los márgenes productivos y la velocidad con que los productos son publicados.

Con el avance de la informática y la notoria preferencia el mundo de hoy por el uso de recursos multimedia, aumenta la construcción de bases de datos en Internet en las que se acumula grandes cantidades de información que encierran formas de conocimiento en sí mismas. Descubrir dicho conocimiento se ha convertido en un interés de la ciencia, pero

no es posible acceder a este objetivo con formas tradicionales de investigación, pues el volumen de datos a los que se tiene acceso supera la capacidad humana, por lo cual se recurre a mecanismos propios de las ciencias de la información, de forma que, por medio de inteligencia artificial, se descubre el conocimiento presente en tales datos.

A esta metodología se le ha denominado como Descubrimiento de Conocimiento en Bases de Datos, más conocida en la literatura internacional como KDD por ser la sigla en inglés de *Knowledge Discovery Database*. De acuerdo con esta perspectiva, los datos *per se* no ofrecen elementos de juicio frente a la información que aportan, es decir, no tienen beneficios directos, sino que la importancia de los mismos estriba en la capacidad de extraer información que permita tomar decisiones y generar comprensión el fenómeno que gobierna la fuente de los datos (Riquelme, Ruiz & Gilbert, 2006).

Esto supone recurrir a estrategias de análisis de datos no convencionales, que por ende superan los métodos estadísticos habituales o cuanto menos les complementan, para ello el KDD implica el cumplimiento de procesos novedosos de identificación de patrones válidos que aporten a la comprensión de los datos evaluados (Gorbea-Portal, 2013), es decir, que permitan obtener el conocimiento útil de los datos (Riquelme et al., 2006).

Esos procesos de los que hablamos, consisten en tareas ordenadas que facilitan el aprovechamiento de los datos en el propósito extractivo de información. Han sido definidos originalmente por Fayyard, Piatetsky-Shapiro y Smyth (1996) y parten de la remoción de datos que generan ruido (*limpieza)* con el fin de evitar inconsistencias en el análisis; realización de múltiples combinaciones (de mayor a menor)

de los datos fuentes (*integración*); consolidación de los datos en formas más apropiadas que su versión original ruidosa (*transformación*); extracción de patrones de datos a partir del uso de métodos inteligentes *(minería);* identificación de los patrones que representan de mejor forma el conocimiento extractado de los datos *(evaluación de patrones);* empleo de técnicas de visualización y representación del conocimiento minado *(representación del conocimiento)*.

Como puede notarse, la Minería de Datos es un procedimiento que forma parte de la metodología KDD. Su desarrollo supone el uso de algoritmos computacionales para la extracción de modelos o patrones de los datos a partir de los cuales es posible predecir tendencias y comportamientos, coadyuvando con ello a la toma de decisiones basadas en el conocimiento derivado de la información minada (Palomo, 2010). Para esto, la minería aprovecha los aportes de diversas áreas como la estadística, la computación gráfica, la inteligencia artificial, entre otras, lo que le hace un campo interdisciplinar cuyo objetivo es predecir e identificar relaciones entre los datos (Mitra & Acharya, 2003).

El uso de algoritmos sofisticados es el mecanismo sobresaliente de la Minería en la tarea extractiva, con ellos se puede descubrir patrones descriptivos o asociativos de los datos o construir clasificaciones de datos nuevos a partir los previamente disponibles (Riquelme et al., 2006). Estos algoritmos se dividen en *supervisados*, que facilitan los análisis predictivos, y *no supervisados*, que se basan en el descubrimiento de conocimiento. A cada tipo de algoritmo aplica una serie de técnicas disponibles para el proceso de minería según sea el objetivo de la investigación. En la Tabla 7.1 se relacionan las principales técnicas de análisis en el minado de datos.

Tabla 7.1. *Principales técnicas de minería de datos según el tipo de algoritmo.*

Técnicas Supervisadas	Descripción
Árboles de decisión	Consiste en la construcción de diagramas en los cuales se construyen reglas para la clasificación de un conjunto de datos. Se emplean para categorizar una serie de condiciones que se presentan de forma sucesiva, de cara a la resolución de un problema.
Redes neuronales	Es una técnica de inteligencia artificial que permite identificar categorías comunes en los datos; puede detectar patrones complejos simulando el funcionamiento del sistema de interconexión neuronal.
Regresión	Busca descubrir relaciones entre variables mediante técnicas lineales y no lineales.
Series temporales	Identifica patrones entre una gran cantidad de datos. Se basa en la extracción de información y la definición tendencias a lo largo del tiempo, por lo que se emplea a partir del comportamiento histórico de los datos.
Técnicas no supervisados	**Descripción**
Reglas de asociación	Genera o extrae reglas de los datos a partir de las cuales descubren relaciones de asociación y dependencias funcionales.
Segmentación	Clasifica un dato dentro de clases definidas.
Agrupamiento	Agrupa los datos de acuerdo con la similitud que haya entre ellos. Las agrupaciones (cluster) de registros generadas son similares entre sí y a la vez son diferentes de otras agrupaciones.
Patrones secuenciales	Identifica patrones similares en un conjunto de datos durante un periodo determinado. Permite la construcción de varias secuencias de patrones con el fin de estudiar tendencias hacia el futuro.

Fuente: Elaborado a partir de Dueñas-Reyes, 2009; Riquelme et al., 2006.

La propuesta desarrollada en este capítulo se basa en el estudio de dos tipos de técnicas, una no supervisada (generación de reglas de asociación), y una supervisada (árboles de clasificación y regresión), con el fin de identificar sus aplicaciones y

utilidades dentro del estudio cienciométrico. Para este fin, a continuación se describe cada técnica, el algoritmo correspondiente y su proceso metodológico, además, aplicaremos el análisis de minería en la evaluación de dos conjuntos de datos, por un lado, generaremos reglas de asociación frente a la producción científica de grupos de investigación en Ciencias Sociales en Colombia, y en segunda instancia se construirá un árbol de clasificación y regresión para predecir el enfoque metodológico de los artículos publicados por dichos grupos, en función de las propiedades bibliométricas de esos productos.

7.2 Reglas de asociación. Utilidades en el estudio bibliométrico

La técnica dirigida a la generación de reglas de asociación se basa en el uso de algoritmos no supervisados debido a que el análisis se cumple sin que se conozcan relaciones de antemano con las cuales se haga un contraste de los resultados, en su lugar se cumple un análisis de la significación estadística de las reglas obtenidas o generadas (García & Álvarez, 2010).

Los algoritmos de asociación permiten descubrir relaciones de forma automática entre los datos contenidos en una base. El procedimiento consiste en identificar reglas que definen las relaciones o asociaciones en un conjunto frecuente de datos (García & Álvarez, 2010). Para ilustrar la lógica que orienta el procedimiento y comprender el sentido de *relaciones en un conjunto frecuente de datos*, nos valdremos de la simpleza descriptiva de Brossette et al., (1998).

Supongamos que en un supermercado se encuentra el reporte de todos los productos que pasan por la caja registradora. Entonces, los productos que son incluidos en las cestas de compra representan registros en esa base de datos. Si en la

base se identifica el conjunto frecuente (pan, queso, leche), probablemente se deba a que durante un solo día esos tres registros (queso, pan, leche), se encuentran juntos en muchas cestas de compra que han pasado por la caja. Ahora bien, como ese conjunto frecuente está dado por la relación de tres registros, entonces también se conforman los conjuntos frecuentes (pan) (queso) (leche), (pan, leche), (leche, queso) y (pan, queso).

Según definen Tan, Steinbach y Kumar (2006), para identificar reglas de asociación tenemos un conjunto de todos los elementos (*items*) de una base de datos $I\{i_1, i_2, ..., i_d\}$, y una tabla de todas las transacciones $T = \{t_1, t_2, ..., t_n\}$. La anchura (*width*) de la transacción es definida por el número de elementos que contiene, de este modo, se dice que una transacción t_j contiene un conjunto de elementos X si X es un subconjunto de t_j.

Podemos definir una regla de asociación como una expresión implicativa de la forma $X \longrightarrow Y$, siendo X e Y conjuntos de datos frecuentes. Para efectos del reglaje X es antecedente e Y consecuente (Belamate, Cassani & Ricci, 2016). La fuerza de la asociación puede ser medida de acuerdo con su soporte (*Support*) y su confidencia (*confidence*). El Soporte determina cómo a menudo una regla es aplicable a un conjunto de datos, por ende, constituye un índice de generación de las combinaciones entre los elementos.

$$s(X \longrightarrow Y) = \frac{\sigma (X \cup Y)}{N}$$

Esta medida es de mucha importancia puesto que una regla con soporte bajo puede estar dada por efecto del azar, de hecho, el bajo soporte implica relaciones de poco interés para la investigación, de forma que con frecuencia es empleado como forma de descartar reglas poco interesantes (Tan et al., 2006).

Por su parte, la confidencia determina con qué frecuencia un elemento en Y aparece contenido en las transacciones de X (Tan et al., 2006). Es decir, es una métrica de la efectividad de la regla (Belamate et al., 2016), o un índice de generación de reglas.

$$c\ (X \longrightarrow Y) = \frac{\sigma\ (X \cup Y)}{\sigma\ (X)}$$

Esta métrica constituye el cálculo del nivel de confiabilidad de la inferencia hecha por una regla. La confiabilidad para una regla $X \longrightarrow Y$ será elevada, cuanto más probable sea que Y esté presente en las transacciones de X.

Adicionalmente, se aplica la métrica denominada elevación (*lift*), que calcula la relación entre X e Y (realmente se entiende como la confianza de la regla dividida por el soporte del consecuente de la regla). Esta métrica obtiene valores iguales, mayores o menores que 1, con los que se identifica la forma de la relación: si lift $>1=$ X e Y correlacionan positivamente; si lift $<1=$ X e Y correlacionan negativamente; si lift $= 1=$ X e Y son independientes. Entre mayor sea el valor, más alta es la probabilidad de que la existencia de una transacción $X \longrightarrow Y$ no sea aleatoria.

$$li\ (X \longrightarrow Y) = s\ (X \longrightarrow Y)\ /s\ (Y)$$

En una base de datos voluminosa resulta altamente complejo identificar todas las posibles asociaciones que surgen entre los ítems (digamos variables), el número de relaciones puede ser tan elevado que incluso los recursos computacionales podrían verse limitados. La solución a este problema se basa en el uso de algoritmos de poda que reducen el reglaje a aquellas relaciones que son estadísticamente significativas. En otras palabras, el procedimiento de poda nos dice si un conjunto es infrecuente, infiriendo que todos los conjuntos en los que ese primer conjunto se encuentre, también lo serán, por

ende, se desatiende la información que los mismos presentan (Brossette et al., 1998).

Esta es, precisamente, la lógica empleada en el *algoritmo a priori*, el cual fue uno de los primeros algoritmos desarrollados para la minería de reglas de asociación y constituye una medida de ayuda de soporte para reducir el número de elementos considerados dentro de un conjunto de elementos frecuentes (*Support-based pruning*). A continuación, aplicaremos estos principios en un conjunto real de datos cienciométricos

7.3 Ejemplo de estudio 1. Reglas de asociación (*a priori*) aplicadas a la producción de grupos colombianos de investigación en Ciencias Sociales

Los datos registrados en este capítulo se derivan del trabajo desarrollado por Ávila-Toscano, Romero-Pérez, Saavedra Guajardo & Marenco-Escuderos (2018), sobre la evaluación cienciométrica del campo de estudios en ciencias sociales en Colombia, investigación financiada por el Departamento Administrativo de Ciencia y Tecnología de este país.

Contamos con una base de datos de los productos científicos de 168 grupos de investigación en Ciencias Sociales, los cuales han sido clasificados por el Sistema Nacional de Ciencia, Tecnología e Innovación de Colombia (SCIENTI-Col). Los grupos están divididos por áreas del conocimiento: Psicología (n=20), Derecho (n=35), Educación (n=20), Sociología (n=59), Ciencias Políticas (n=8), Periodismo (n=17) y Otras Ciencias Sociales (n=9). Además, la base describe la clasificación de estos grupos en A1, A y B, es decir, la categoría que obtienen dentro de SCIENTI-Col según su nivel de producción, índices de cooperación, niveles de integración, entre otros elementos, siendo A1 la máxima categoría.

La base de datos incluye toda la producción generada entre 2006 y 2015, y registrada en la plataforma tecnológica de acceso público donde se ingresa el historial productivo de los grupos de investigación en Colombia (GrupLAC). Nuestro objetivo se basó en identificar reglas de asociación definidas para la producción obtenida por los grupos y sus atributos como clasificación y área del conocimiento. Para ello, los elementos incluidos en el análisis se describen en la Tabla 7.2.

Tabla 7.2. *Elementos considerados en el análisis de minería de reglas de asociación.*

Elemento (ítems)	Descripción
Clasificación del grupo	A1, A, B
Área de conocimiento	Psicología, Derecho, Educación, Sociología, Ciencias Políticas, Periodismo, Otras Ciencias Sociales
Artículos incluidos en WoS/Scopus	Publicaciones en revistas incluidas en las bases de WoS o Scopus
Artículos incluidos tipo C/D	Publicaciones tipo artículo en índices regionales o internacionales distintos a WoS o Scopus. Reciben el nombre de "tipo C/D" por su denominación dentro del modelo de medición del SCIENTI-Col.
Libros de investigación	Libros que publican resultados de investigación
Capítulos de libro investigación	Capítulos incluidos en libros que publican resultados de investigación
Libros de divulgación	Libros que publican contenidos académicos no derivados directamente de un proyecto de investigación
Capítulos de libros de divulgación	Capítulos de libros que publican contenidos académicos no derivados directamente de un proyecto de investigación
Tesis de maestría	Tesis de maestría defendidas y aprobadas
Tesis de doctorado	Tesis doctorales defendidas y aprobadas
Producción endógena	Artículos WoS/Scopus o tipo C/D publicados en revistas publicadas por las instituciones a las que se adscriben los grupos.

Fuente: elaboración propia.

Los datos fueron analizados con el software de Minería WEKA (*Waikato Environment for Knowledge Analysis*) versión 3.8.1, desarrollado por investigadores de la Universidad de Waikato en Hamilton, Nueva Zelanda. El proceso de creación de reglas se cumplió con el algoritmo *a priori*, tras la discretización de los datos, puesto que este algoritmo trabaja con datos simbólicos.

Inicialmente, se eliminaron los productos de divulgación (libros y capítulos de libro) porque en el análisis inicial no ofrecieron resultados significativos. De todos los conjuntos de reglas generadas, en la Tabla 7.3 se describen las más importantes. Es necesario aclarar que el software WEKA ofrece cuatro métricas, las ya conocidas Confianza (Conf), Elevación (Lift), y los indicadores de Apalancamiento (Leverage=Lev) y Convicción (Conviction-Conv). El apalancamiento mide la proporción de casos de X e Y por encima de lo esperado, si X e Y son independientes entre sí. La convicción determina el efecto del incumplimiento del consecuente de la regla.

Tabla 7.3. *Reglas de asociación obtenidas.*

Regla	Conf	Lift	Lev	Conv
1. [WoS_SCOPUS=Baja C_D=Baja] ==> [Endogeno_total=Con endógenos]	0.98	1.22	0.06	5.3
2. [WoS_SCOPUS=Baja C_D=Baja Clasificación=B] ==> [Endogeno_total=Con endógenos]	0.97	1.21	0.05	3.73
3. [Área=Educación C_D=Baja Cap_Lib_Inv=Baja Tesis_PhD=Sin tesis] ==> [WoS_SCOPUS=Baja]	0.96	2.03	0.07	6.09
4. [Cap_Lib_Inv=Si] ==> [Libros_Inv=Si]	0.96	1.01	0.01	1.02
5. [Cap_Lib_Inv=Baja Tesis_Mg=Baja Clasificación=B] ==> [Tesis_PhD=Sin tesis]	0.94	1.68	0.07	4.7
6. [Endógeno_C_D=Con endógenos Cap_Lib_Inv=Baja Tesis_Mg=Baja] ==> [Endógeno_Scopus=Con endógenos]	0.92	1.14	0.03	1.82

Fuente: elaboración propia.

De acuerdo con las reglas se identifica un papel relevante del elemento *producción Endógena*, presente en tres de las reglas que hemos destacado. Según el minado, 98 % de los grupos de investigación con poca producción de artículos, realiza producciones endógenas; además 97 % de los grupos clasificados en categoría B, con la baja producción ya descrita tanto en artículos WoS/Scopus como Tipo C/D, genera producción endógena; por último, 92 % de los grupos que tiene artículos endógenos tipo C/D, pocos capítulos de libros publicados y pocas tesis de maestría asesoradas, también genera artículos WoS/Scopus en condiciones endogámicas.

Entre otros resultados tenemos que los grupos del área de Educación con bajos indicadores de artículos tipo C/D, tesis de doctorado y capítulos de libros, también producen pocos artículos WoS/Scopus.

Observemos además que los valores de *lift* son superiores a 1, por lo cual se asume que los elementos se asocian de forma positiva y además ello es indicador de que la regla hacia el futuro tiene más probabilidades de que se repita. Sin embargo, en el caso de la regla número 4, dada la cercanía de *lift* a 1 y el apalancamiento a cero (0), nos obliga a ser conservadores optando por la omisión de la misma dentro del análisis extractivo.

7.4 Árboles de clasificación y regresión. El Algoritmo CART

Los árboles de clasificación y regresión se basan en un proceso de aprendizaje inductivo mediante la partición binaria recursiva para obtener segmentos de datos empleando un conjunto de variables o criterios de clasificación. En otras palabras, permiten asignar elementos (ítems, variables,

individuos) de una muestra a diferentes categorías o grupos a partir de una variable determinada (Richard´s et al., 2008; Trujillano et al., 2008; Wu et al., 2008).

En este apartado no enfocaremos en el algoritmo CART desarrollado por Breiman, Friedman, Olshen y Stone (1984), el cual es considerado como uno de los más importantes avances metodológicos aplicado a la inteligencia artificial, el lenguaje de máquinas, la minería de datos y el análisis de datos no paramétrico (Wu et al., 2008).

La sigla CART proviene del inglés *Classification and Regression Trees*. Se trata de un procedimiento de minería de datos basado en la creación de árboles a partir de los cuales se construyen clasificaciones de los datos observados y se desarrollan procedimientos de regresión (Breiman et al., 1984). CART es un procedimiento no paramétrico de partición recursiva binaria, que permite trabajar con conjuntos de datos categóricos y continuos (Steinberg, 2009); el algoritmo emplea una técnica que añade paso a paso términos y realiza una función de poda (*pruning funtion*) a través de la selección de todas las variables con mayor nivel de importancia dentro de un conjunto de datos, las cuales resultan útiles para correr procedimientos de regresión (Low, Ng, Kabir, Koh & Sinnasamy, 2014).

La construcción del árbol comienza con un nodo inicial al que se le denomina *nodo padre*, el mismo incluye todos los valores contenidos en la base de datos. El nodo padre se divide a partir de una función de partición (*splitting function*) por medio de la cual se identifica la variable más adecuada para partirlo en dos nodos hijos que se dirigen a la derecha y la izquierda (Daud, Ahmad, Malik & Che, 2015).

Siendo t_p un nodo padre, t_l un nodo hijo a la izquierda y t_r un nodo hijo a la derecha del nodo padre t_p; x_j una variable *j*,

y x_j^R, la variable x_j con mejor valor de partición. La partición del nodo padre en dos nodos hijos, mediante la selección de la mejor variable, tiene que hacerse con la máxima homogeneidad que se define por medio de una función de impureza (*impurity function*) *i* (*t*).

Considerando que la impureza del nodo padre t_p es constante en todas las divisiones posibles $x_j \leq x_j^R$, *j*=1,..., *M*, la máxima homogeneidad de los nodos hijos equivaldrá al cambio de la función de impurezas Δ*i*(t).

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)$$

En el árbol, se utiliza en primer lugar la variable que obtiene mejor nivel de pureza, y sucesivamente se van integrando las demás variables. En cada nodo, CART resuelve el siguiente problema de maximización.

$$\arg \max_{x_j \leq x_j^R,\ j=1,\ldots,m} [i(t_p) - P_l i(t_l) - P_r i(t_r)]$$

Para garantizar que la función de partición (*splitting function*) asegure el mayor nivel de pureza en los nodos hijos, se emplean índices que buscan la optimización de la pureza, el más comúnmente usado es el índice de Gini (*Gini Index*), el cual alcanza un índice de pureza que se considera como máximo y tiene excelente desempeño con datos ruidosos. El Índice de Gini es ampliamente favorecido en la literatura CART (Steinberg, 2009), y utiliza la siguiente función de impureza $i(t)$:

$$i(t) = \sum_{k \neq l} p(k|t)p(l|t)$$

Siendo *k*, *l* 1,...,*K* el índice de la clase; *p* (*k* | *t*) la probabilidad condicional de la clase k siempre que se encuentre en el nodo *t*.

7.4.1 *Usos de los árboles de clasificación y regresión en las ciencias de la información*

Los árboles de clasificación son ampliamente útiles en el estudio de la ciencia, tanto en la identificación de clasificaciones a partir de indicadores bibliométricos como en la identificación de conjuntos de datos relacionados con la innovación industrial o la toma de decisiones empresariales. En gran medida, su utilidad se relaciona con las ventajas que implica su uso dado que los árboles aportan reglas, es decir, extraen conocimiento de conjuntos de datos que permite adoptar decisiones, pueden también predecir la ocurrencia de fenómenos con base en los criterios de clasificación y son sumamente sencillos de entender por cuanto se basan en la reducción de variables ofreciendo resultados con contenido visual muy comprensibles (Berlanga Silvente, Rubio Hurtado & Vilà Baños, 2013).

Incluso personas poco avezadas en el uso de la estadística pueden realizar interpretaciones solventes de los resultados, pues se trata de un procedimiento intuitivo. Adicionalmente, los árboles permiten hacer guardado de las clasificaciones obtenidas y sus reglas y pronósticos (Berlanga Silvente, 2013).

Por su parte, diversos autores han empleado el método CART para identificar procedimientos de mejora en la investigación con miras a lograr avances en el desarrollo y diseño de productos en la industria de los moldes (Yeh, Cheng & Hsiao, 2011); la predicción del grado de asociación de indicadores bibliométricos como citas, factor de impacto y niveles de las revistas, con la colaboración internacional en publicaciones científicas (Low et al., 2014), entre otros procesos.

A continuación, exploraremos la aplicación de CART a un caso real de estudio.

7.5 Ejemplo de estudio 2. Identificación de determinantes bibliométricos del enfoque metodológico de artículos científicos en Ciencias Sociales aplicando el algoritmo CART

La investigación en Ciencias Sociales ha venido teniendo una transición en cuanto a los formatos de publicación empleados, puesto que la forma tradicional de divulgación del conocimiento en estas disciplinas venía siendo el libro, sin embargo, cada vez es más común que los científicos sociales realicen publicaciones en revistas incluidas en índices internacionales como WoS y Scopus.

Otra discusión importante al respecto consiste en los tipos de enfoques metodológicos realizados por los investigadores de estas disciplinas para el desarrollo de sus trabajos. Es bien sabido que la mayoría de las revistas internacionales de alto impacto difunden resultados de investigaciones reproducibles, de forma que el enfoque común es el cuantitativo. En otras disciplinas como las ciencias básicas este no es objeto común de análisis, pues producen investigaciones predominantemente cuantitativas, sin embargo, en Ciencias Sociales, por la naturaleza de su objeto de estudio es común el uso de perspectivas centradas en la comprensión de fenómenos más que en su predicción o explicación.

Los resultados que presentamos en este caso ilustrativo se basan en la identificación de variables que predicen el enfoque metodológico de los artículos científicos en Ciencias Sociales publicados por investigadores colombianos. La base de datos consistió en 2992 artículos de siete disciplinas sociales publicados por autores adscritos a los grupos de investigación descritos en el primer ejemplo (Véase numeral 7.3 de este capítulo).

El análisis se cumplió con el algoritmo CART como método de crecimiento del árbol, determinando el *enfoque metodológico* como variable dependiente, mientras que las independientes o predictoras fueron la disciplina o área del conocimiento, la producción endógena, la categoría del artículo, el tipo de firma, el tipo de colaboración y el tipo de artículo. La descripción de todas las categorías se encuentra en la Tabla 7.4.

Tabla 7.4. *Descripción de las variables incluidas en el árbol.*

Variables	Categorías
Enfoque metodológico	Cuantitativo, Cualitativo, Mixto
Área del conocimiento	Psicología, Derecho, Educación, Sociología, Ciencias Políticas, Periodismo, Otras Ciencias Sociales
Producción endógena	Si, No
Categoría del artículo	WoS/Scopus, Tipo C/D
Tipo de firma	Único autor, Coautoría
Tipo de colaboración	Intrainstitucional, Interinstitucional, Internacional
Tipo de artículo	Reflexión/Teórico, Investigación aplicada/instrumental, Investigación básica

Fuente: elaboración propia.

El análisis se cumplió con el software SPSS versión 19. Inicialmente se generó un árbol de 19 nodos, 10 de ellos terminales, con una profundidad de 5, el cual incluyó todas las variables predictoras. La estimación del *Riesgo* fue de .174 (Error=.007) lo cual indica que los enfoques metodológicos de los artículos se predijeron con un 84 % de certeza. Se trató de un árbol con alta "frondosidad" de sus ramas. En este caso la visualización del árbol corresponde a la ilustración de la Figura 7.1.

Con el fin de reducir el sobreajuste de un árbol frondoso como el obtenido, se realizó el procedimiento de poda, lo que permite hacer el recorte de las ramas una vez el árbol haya llegado a su máxima profundidad, de esta manera el árbol se

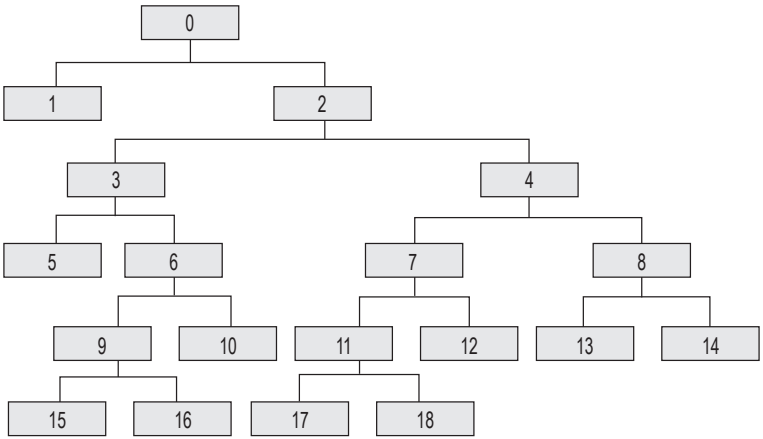

Figura 7.1. *Representación del árbol obtenido en la fase inicial de análisis aplicando el método de crecimiento CART.*

poda hasta su profundidad más pequeña con niveles aceptables de riesgo. El resultado fue un árbol con profundidad de 4, con 9 nodos hijos, 5 de ellos terminales. La estimación del riesgo fue ligeramente más alta que en la primera fase (Riesgo=.161, Error=.007), sin embargo, el árbol sostuvo una predicción de 84 % de los casos incluidos, lo que se toma como una medida buena. Los resultados indican que el árbol ofrece buenos niveles de clasificación tanto para los artículos de enfoque cualitativo (82.5 %), como para aquellos con enfoque cuantitativo (86.5 %) (Tabla 7.5).

Tabla 7.5. *Resumen de clasificación.*

Observado	Pronosticado		
	Cuantitativo	Cualitativo	% correcto
Cuantitativo	888	138	86.5%
Cualitativo	345	1621	82.5%
% global	41.2%	58.8%	83.9%

Fuente: elaboración propia.

Los resultados definitivos del árbol se describen en la Figura 7.2 y en la Tabla 7.6; si bien ambos recursos contienen la misma información, el ánimo de presentarlos no está en ser redundantes sino en mostrar los dos tipos de salidas que se pueden aprovechar para la lectura del árbol.

Como se aprecia en ambos recursos (Figura y Tabla), son cuatro las variables predictoras esenciales: el Tipo de artículo, con un nivel de importancia de 100 % (Importancia=.145); Tipo de firma con una importancia dentro del modelo de 81.8 % (Importancia=.119); la Disciplina, con 69.8 % (Importancia=.101); y la Categoría del artículo con 3.7 % (Importancia=.046).

Un artículo teórico tiene 34.5% de probabilidades de ser de enfoque cualitativo, mientras que los artículos de investigación aplicada, instrumental o básica tienen 65.5 % de probabilidades de ser cuantitativos. Del mismo modo, los artículos que difunden resultados de investigación aplicada tienen 31 % de probabilidad de ser de Otras Ciencias Sociales o Psicología, mientras que en las demás disciplinas se predice la producción de enfoque cualitativo en 34.5 %. En estas disciplinas, es decir, aquellas distintas a Otras Ciencias Sociales y Psicología, la producción de enfoque cualitativo es la más probable tanto en artículos incluidos en WoS y Scopus como en otros índices internacionales o regionales. Finalmente, la producción incluida en WoS y Scopus generada en coautoría es de enfoque cuantitativo (10.2 % de probabilidad), mientras que los trabajos cualitativos suelen ser de único autor (4.2 % de probabilidad).

7.6 Discusión

Como hemos descrito, el principio *a priori* conduce a la generación de reglas de asociación a partir del teorema "si

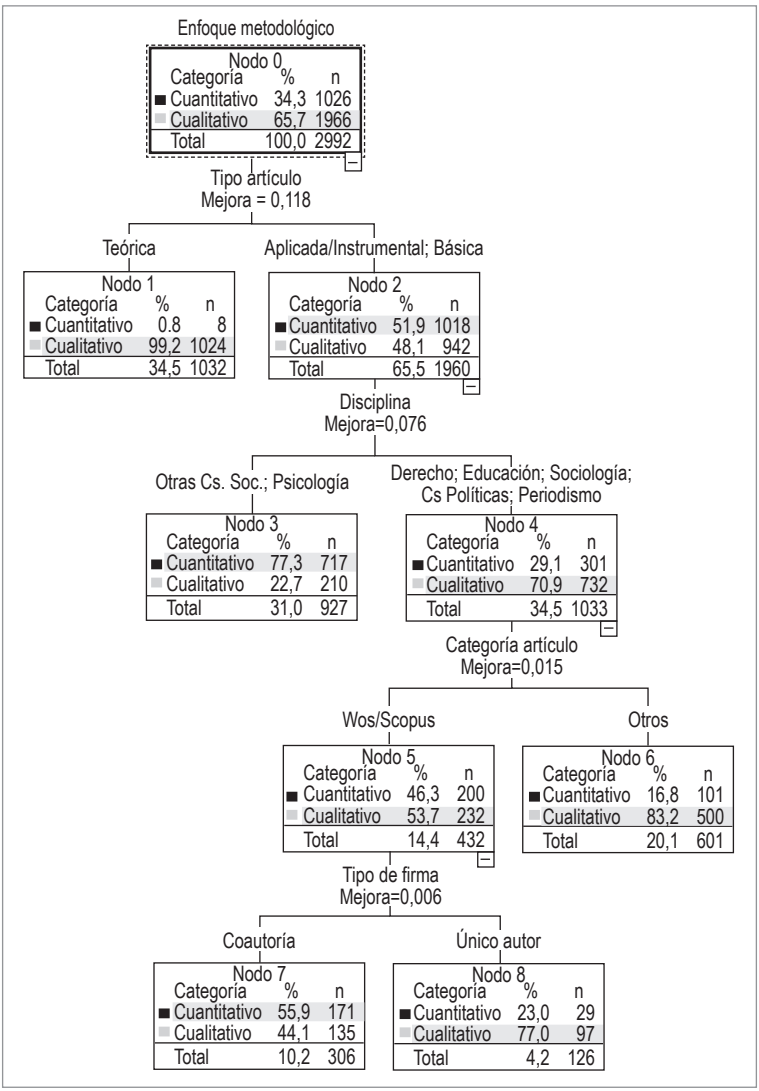

Figura 7.2. *Árbol de clasificación y regresión.*

Tabla 7.6. *Salida tabulada del Árbol de clasificación y regresión.*

Nodo	Cuantitativo		Cualitativo		Total		Categoría pronosticada	Nodo padre	Variable independiente primaria		
	N	%	N	%	N	%			Variable	Mejora	Valores de segmentación
0	1026	34.3	1966	65.7	2992	100.0	Cualitativo				
1	8	.8	1024	99.2	1032	34.5	Cualitativo	0	Tipo artículo	.118	Teórica
2	1018	51.9	942	48.1	1960	65.5	Cuantitativo	0	Tipo artículo	.118	Aplicada/Instrumental; Básica
3	717	77.3	210	22.7	927	31.0	Cuantitativo	2	Disciplina	.076	OSC; Psicología;
4	301	29.1	732	70.9	1033	34.5	Cualitativo	2	Disciplina	.076	Der.; Educ.; Soci.; Polít:, Period.
5	200	46.3	232	53.7	432	14.4	Cualitativo	4	Categoría artículo	.015	Wos/Scopus
6	101	16.8	500	83.2	601	20.1	Cualitativo	4	Categoría artículo	.015	Otros
7	171	55.9	135	44.1	306	10.2	Cuantitativo	5	Tipo de firma	.006	Coautoría
8	29	23.0	97	77.0	126	4.2	Cualitativo	5	Tipo de firma	.006	único autor

OSC=Otras Ciencias Sociales

Fuente: elaboración propia.

un conjunto de elementos es frecuente, entonces todos los subconjuntos que los contengan también lo serán". Entonces, si el conjunto {A, B, C} es frecuente, las transacciones que contengan estos elementos serán frecuentes. La utilización de este método facilita la detección de asociaciones (no causales) entre variables dentro de un conjunto voluminoso de datos, como los que usualmente se emplean dentro de los estudios bibliométricos.

Como hemos visto, la aplicación del algoritmo *a priori* para la detección de reglas de asociación constituye una alternativa valiosa en el campo del estudio de la producción científica en tanto facilita la construcción de reglajes que ayudan a identificar conjuntos frecuentes de datos partiendo de la idea de que no se cuenta con una presunción previa de posibles relaciones esperadas.

Por otro lado, los árboles de clasificación y regresión permiten la asignación de datos a diferentes nodos que representan nuevas agrupaciones a partir de las cuales se realizan procedimientos de regresión que definen la probabilidad de pertenencia a los grupos identificados según se cumplan ciertos supuestos. En los ejemplos presentados hemos mostrado la utilidad de ambos procedimientos ante objetivos bibliométricos reales, reconociendo así las bondades de estos procedimientos de análisis extractivo.

El uso de métodos alternativos a los indicadores bibliométricos tradicionales (Véase Capítulo 4) permite proponer estrategias de análisis que complementan la detección de indicadores de producción. En el caso de los algoritmos descritos se facilita trabajar con datos cuantitativos y discretizados lo que constituye una ventaja importante a la hora de escoger las herramientas metodológicas pertinentes. Estos métodos permiten superar el carácter estático de gran parte de los datos bibliométricos puesto que incluso facilitan (en el caso de los árboles) la generación de

marcadores de participación que leídos de la forma adecuada pueden entenderse como elementos clasificatorios predictivos de una determina condición. En el caso de los ejemplos presentados, los análisis contribuyen, por un lado, a reconocer cómo la clasificación de un grupo de investigación se asocia con la producción o no de determinado tipo de producto científico, y por otra parte, permiten reconocer las características de definen la selección de un determinado enfoque científico a la hora de producir investigaciones en ciencias sociales.

Referencias

Ávila-Toscano, Romero-Pérez, Saavedra Guajardo & Marenco-Escuderos (2018). *Cienciometría del campo de estudios en Ciencias Sociales en Colombia (2006-2015): Análisis de la producción científica, redes de coautoría, cooperación institucional y posicionamiento de grupos de investigación*. Departamento Administrativo de Ciencia, Tecnología e Innovación COLCIENCIAS. Programa Nacional de Ciencia, Tecnología e Innovación en Ciencias Humanas Sociales y Educación. COD. SIGP: 01800-740-54754

Belamate, D., Cassani, M. & Ricci, C. (2016). *Aplicación de reglas de asociación para la detección de patrones de comportamiento en sistema académico universitario***.** Universidad Tecnológica Nacional. Argentina. Disponible en: http://cytal.frvm.utn.edu.ar/q/tf/7/62

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984) Classification and Regression Trees. *Wadsworth International Group*. Belmont, CA, EEUU.

Brossette, S., Sprague, A., Hardin, J. M., Waites, K. B., Jones, W. T. & Moser, S. A. (1998). Association Rules and Data Mining in Hospital Infection Control and Public

Health Surveillance. *Journal of the American Medical Informatics Association*, *5*(4), 373–381. DOI: https://doi.org/10.1136/jamia.1998.0050373

Daud, A., Ahmad, M., Malik, M.S.I & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, *102*(2), 1687-1711. DOI 10.1007/s11192-014-1455-8

Dueñas-Reyes, M. (2009). Minería de datos espaciales en búsqueda de la verdadera información. *Ingeniería y Universidad*, *13*(1), 137-156.

Fayyard, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, (*17*)3, 37-54.

Franceschet, M. (2010). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, *4*, 55–63.

García, M. & Álvarez, A. (2010). *Análisis de Datos en WEKA – Pruebas de Selectividad*. Universidad Carlos II de Madrid. Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

Gorbea-Portal, S. (2013). Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimiento. *Perspectivas em Gestão & Conhecimento*, *João Pessoa*, (*3*)1, 13-27.

Letina, S. (2016). Network and actor attribute effects on the performance of researchers in two fields of social science in a small peripheral community. *Journal of Informetrics*, *10*, 571–595.

Low, Y., Ng, K. H., Kabir., Koh, M. A. & Sinnasamy, J. (2014). Trend and impact of international collaboration in clinical medicine papers published in Malaysia. *Scientometrics*, *98*(2), 1521–1533 DOI 10.1007/s11192-013-1121-6

Mitra, S. & Acharya, T. (2003). Data mining: multimedia, soft computing and bioinformatics. New Jersey, USA: John Wiley & Sons.

Palomo, O. (2010). *Minería de datos*. Universidad Carlos III de Madrid.

Richard´s, M., Solanas, A., Ledesma, R., Introzzi, I. & López, M., (2008). Técnicas estadísticas de clasificación: un estudio comparativo y aplicado. *Psicothema*, *20*(4), 863-871.

Riquelme, J., Ruiz, R. & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial*, *Revista Iberoamericana de Inteligencia Artificial*, *29*, 11-18.

Steinberg, D. (2009). CART: Classification and Regression Trees. In: X, Wu. & V. Kumar (Eds.). *The top ten algorithms in data minning*. (pp.179-201). Chapman & Hall/CRC Taylor & Francis Group. Boca Raton, FL.: USA.

Tan, P-N., Steinbach, M. & Kumar, V. (2006). *Introduction to data mining*. 2nd Ed. USA: Pearson.

Trujillano, J., Sarria-Santamera, A., Esquerda, A., Badia, M., Palma, M. & March, J. (2008). Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. *Gaceta Sanitaria*, *22*(1), 65-72.

Verleysena, F.T. & Engels, T.C.E. (2014). Barycenter representation of book publishinginternationalization in the Social Sciences and Humanities *Journal of Informetrics*, *8*, 234–240.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., (…) Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge Informatics Systems*, *14*, 1-37. DOI: 10.1007/s10115-007-0114-2

Yeh, D., Cheng, C. & Hsiao, S. (2011). *Journal of Intelligent Manufacturing*, *22*(4), 585-595. doi>10.1007/s10845-009-0321-7