UNIVERSIDAD DE LA HABANA

INSTITUTO SUPERIOR DE TECNOLOGÍAS Y CIENCIAS APLICADAS

PROGRAMA DOCTORAL DE GESTIÓN DE LA CIENCIA, LA TECNOLOGÍA, LA INNOVACIÓN Y EL MEDIOAMBIENTE

Metodología para el estudio de Datos Bibliográficos con el empleo de la Minería de Datos en la Biblioteca de Ciencias y Técnicas

Tesis presentada en opción al grado científico de Doctor en

Ciencias Técnicas

Autor: **Ing. Esther Marina Ruiz Lobaina**

Tutores: **Dr.C. Pedro Lázaro Romero Suárez**

Dr.C. Juan Pedro Febles Rodríguez

La Habana
2023

Agradecimientos

Gracias Yeshúa, por darme esta resistencia tan grande y soportar todo este tiempo sin claudicar.

Muchas gracias Yeshúa por poner en mi camino al tutor y amigo Dr.C. Pedro Lázaro Romero Suárez, porque sin él nunca hubiera sido posible alcanzar esta meta, él es mi motor. Gracias Yeshúa por el Dr.C. Juan Pedro Febles Rodríguez, que es mi cotutor y profesor con quien inicié este sueño.

A todos los que me han ayudado de forma directa e indirecta con sus críticas y elogios, aumentando mis fuerzas.

Gracias!

Dedicatoria

Esta investigación está dedicada a aquellas personas que trabajan con la información y que están interesadas en descubrir el conocimiento oculto en ella, con el objetivo de crear o mejorar el resultado de su trabajo en bien de la humanidad.

Síntesis

Esta investigación tiene como objetivo desarrollar una metodología para mejorar la calidad de la información de las bases de datos bibliográficas con la aplicación de algunas de las técnicas de minería de datos más conocidas e impulsar la gestión de la información.

Durante el desarrollo de la metodología se eliminan los obstáculos fundamentales que han provocado, que estas técnicas de exploración de la información aún no se empleen masivamente en las bibliotecas, a pesar de que ya tienen algunos años de uso en muchos sectores productivos, económicos y sociales en la vida del hombre.

Entre los aportes logrados con la aplicación de esta metodología que utiliza la minería de datos, está el mejoramiento del funcionamiento del sistema gestor bibliotecario y la obtención de los patrones de comportamiento (información no visible a simple vista), con la cual se ha creado un repositorio que sirve como guía para la búsqueda de nueva información en bases de datos internacionales y de patentes, además de la creación de nuevos productos y servicios bibliotecarios que apoyan la difusión selectiva de información, la vigilancia e inteligencia y la prospección.

Palabras Claves: minería de datos, gestión de la información, bases de datos bibliográficos, clusterización, descubrimiento de conocimiento.

INDICE

Anexos, Figuras y Tablas

DSI - Difusión Selectiva de Información.
SOM - Self Organizing Maps
OLAP - Online Analytical Processing
OPAC - Online Public Access Catalog

Introducción

Introducción

El Decreto-Ley No. 7 / 2021 sobre el Sistema De Ciencia, Tecnología e Innovación, Capítulo I establece en el artículo 3 (MINJUS, 2021) que las actividades de ciencia, tecnología e innovación son aquellas actividades sistemáticas que están estrechamente relacionadas con la producción, difusión y utilización del nuevo conocimiento en los diferentes campos de la ciencia y la tecnología con impacto en la economía y la sociedad, comprendiendo las de investigación y desarrollo (I+D), la innovación, los servicios científicos y tecnológicos, las producciones especializadas, las actividades de interface y la transferencia de tecnología.

En el mismo Decreto el Capítulo II De Los Componentes, Objetivos Y Principios del Sistema, Sección Segunda en el Artículo 5 (MINJUS, 2021) define que el Sistema tiene entre sus objetivos algunos aspectos como:

a) Fomentar la generación, asimilación y aplicación de conocimientos y tecnologías;

b) incrementar la investigación y la innovación en el campo de las ciencias sociales y fortalecer su utilización en todos los sectores y niveles de dirección, como herramienta imprescindible para enriquecer el impacto de la actividad de ciencia, tecnología e innovación en la economía y la sociedad cubana;

c) contribuir a la formación de valores y al fortalecimiento de la conciencia nacional;

d) estimular y propiciar el aprendizaje y la innovación en las esferas de la vida económica y social del país en todas las instancias, con el fin de contribuir al desarrollo sostenible;

e) incrementar el aporte de la ciencia, la tecnología y la innovación para el desarrollo económico y social, mediante la integración entre sus actores, en cumplimiento de los requerimientos de la sociedad; e innovadora en la sociedad.

Para implementar estas orientaciones en el ámbito de las bibliotecas, es necesario el estudio de grandes cantidades de información, lo que se facilitaría con el empleo de la minería de datos, ya que es una herramienta de trabajo actual utilizada ampliamente y de manera creciente en la investigación científica y académica.

La minería de datos es una técnica asistida por computadora que se utiliza en los análisis para procesar (Pio, 2018) y explorar grandes conjuntos de datos basados en la estadística y las matemáticas, de este modo las organizaciones pueden descubrir

patrones y relaciones ocultas en sus datos. La minería de datos transforma datos en bruto en conocimiento práctico y extrae un significado o un conocimiento valioso de estos. Las compañías utilizan dicho conocimiento para resolver problemas, analizar las consecuencias en el futuro de decisiones empresariales y aumentar sus márgenes de beneficio (León et al., 2019). Son nuevos conocimientos, que no están implícitos en ningún dato de la información propiamente. Los resultados obtenidos a través del análisis de estos patrones, han tenido mucho éxito en campos como la biotecnología, la industria, los negocios, entre otros, (Acosta, 2019).

Considerando las ventajas que estas nuevas técnicas de análisis en grandes bases de datos proporcionan y reconociendo que las bibliotecas son los espacios con mayor responsabilidad en el desarrollo intelectual de una nación, se puede afirmar de acuerdo con el Decreto-Ley No. 7 / 2021 que la calidad de los servicios bibliotecarios, se convierten en un objetivo que justifica la necesidad de implementar las más modernas herramientas y métodos de digitalización, organización, recuperación y análisis de la información (Fernández et al., 2018) , pero con frecuencia se encuentra que la situación es otra y muchas de ellas aún siguen trabajando en forma tradicional y hasta sin un software para recuperación de información.

En Cuba las bibliotecas adolecen de procesos donde se emplea la minería de datos incorporados a su sistema de gestión por varios factores como son: la tecnología es atrasada, con poco espacio en disco duro, lentos procesadores y poco tamaño de memoria (RAM), que no sirven para procesar grandes bases de datos, además, que la formación académica de los bibliotecarios, en su mayoría, no le permite enfrentar la implementación de estos procesos, que no son habituales en sus contenidos de trabajo (Suárez et al., 2017).

El desarrollo de una metodología para el estudio de datos bibliográficos con el empleo de la minería de datos permite en las condiciones actuales encontrar patrones útiles y cumplir con los objetivos propuestos para el Proyecto de Ciencia e Innovación Tecnológica Red CubaCiencia: Recolector y Directorio Nacional de Ciencia. Para esta investigación y creación de la metodología, se tomó como caso de estudio la base de datos CubaCiencia, que en ese momento constaba con 10492 registros.

La minería de datos ha logrado probar que es un proceso de análisis de datos y extracción de conocimiento, que no tiene similitud con ninguna otra forma de manejo de la información, utiliza sus propias herramientas y contempla varias fases donde en cada una están involucrados diferentes procedimientos y softwares, que permiten obtener los resultados de cada fase en particular, quedando entonces definido que un proceso de descubrimiento de conocimiento contempla la ejecución de todas las fases, empezando desde la fase de selección de los datos a procesar, la limpieza y la estandarización de la información hasta su terminación, que concluye con los informes que muestran los patrones encontrados y validados a través del conocimiento que estos aportan (Marulanda et al., 2017), (Flores, 2021).

Basado en todo el conocimiento anterior y contando que la información que se quiere analizar es información bibliográfica se identificó que *existen dificultades para integrar las técnicas de minería de datos con la gestión de la información bibliográfica en la Biblioteca de Ciencias y Técnicas*, lo que se constituye en el problema científico de esta investigación. Para su solución se formuló como hipótesis que *el desarrollo de una metodología que integre las técnicas de minería de datos a la gestión de la información bibliográfica posibilita la transformación de esta información y la obtención de patrones de comportamiento que contribuyen al mejoramiento de los servicios de información bibliográficos en la Biblioteca de Ciencias y Técnicas.*

A partir del problema científico identificado y la hipótesis formulada para la solución se definió a la información bibliográfica como objeto de estudio y se establecieron como objetivo general y específicos:

Objetivo General

Desarrollar una metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencias y Técnicas.

Objetivos Específicos

1. Determinar el marco teórico referencial relacionado con la aplicación de técnicas de minería de datos en las bibliotecas, las metodologías asociadas a esta disciplina y las razones que no han permitido aplicar de forma masiva, la minería de datos en las bibliotecas.
2. Desarrollar una metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencias y Técnicas.

3. Validar la metodología a través de su aplicación dentro de la Biblioteca de Ciencias y Técnicas, de manera que permita comprobar la hipótesis planteada.

Como resultado de la investigación se consideran **novedades científicas**:

1. La metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencias y Técnicas.
2. El repositorio de información bibliográfica obtenido que recoge los patrones identificados con la aplicación de las técnicas de minería de datos y con cuyo empleo se apoya la difusión selectiva de información (DSI), la vigilancia e inteligencia y la prospección.

Los **beneficios internos** a la biblioteca son los siguientes:

1. El perfeccionamiento del proceso de recuperación de la información en el sistema gestor bibliotecario.
2. El incremento de la calidad en la oferta de productos y servicios que ofrece la biblioteca.
3. La realización eficiente de la gestión de la información.

Los **beneficios sociales** de esta investigación son:

1. Una biblioteca competente y preparada, para brindar un servicio de calidad comparable con las bibliotecas más avanzadas del mundo.
2. Servicios apoyados en el nuevo conocimiento que aportan los patrones encontrados a través de los procesos de minería de datos.
3. Un repositorio para la gestión de la información que sirva como fuente formal de un observatorio dedicado a la vigilancia e inteligencia y prospección del conocimiento y el fortalecimiento de la DSI para los usuarios.

Los **beneficios económicos** son:

1. Se obtiene un ahorro por concepto de compra de herramientas digitales para todo el proceso de minería de datos, porque se utilizan herramientas libres.
2. Se aplican algoritmos que pueden ejecutarse en las computadoras que utiliza el personal de la biblioteca para su trabajo diario, lo que constituye un ahorro en compra de nuevo hardware.

3. Se tiene un ahorro por concepto de salarios al ser el propio personal de la biblioteca quien ejecute los algoritmos y analice los resultados.

En el desarrollo de la investigación se utilizaron un conjunto de técnicas y métodos de investigación científica los cuales se resumen a continuación:

Enfoque sistémico: para concebir el proyecto investigativo, instrumentos y herramientas evaluativas, análisis y sistematización de resultados, (Schmukler, 2017).

Método hipotético-deductivo: en la elaboración de la hipótesis de la investigación y en su validación, (Chagoya, 2023).

Métodos dialéctico e histórico-lógico: para el estudio de los aportes de otros investigadores; de metodologías y guías de evaluación y autoevaluación, (Chagoya, 2023).

Método analítico-sintético: para descomponer el problema de investigación en elementos por separado y profundizar en el estudio de cada uno de ellos, para luego sintetizarlos en la elaboración de la propuesta, (Hernández et al., 2014).

La observación científica y la medición mixta (cuantitativa y cualitativa) debido a que se analizan todos los factores de forma cualitativa y por datos en su forma cuantitativa según corresponda, (Hernández et al., 2014).

La tesis se estructura con una introducción, tres capítulos, las conclusiones, recomendaciones, referencias bibliográficas y anexos.

En el primer capítulo se presenta el marco teórico referencial relacionado con la minería de datos y su empleo en las bibliotecas, el segundo capítulo describe la metodología desarrollada con las herramientas de minería de datos empleadas y en el tercer capítulo se muestran los resultados obtenidos de la aplicación de la metodología en la Biblioteca de Ciencias y Técnicas, así como su análisis y validación.

Capítulo I

Capítulo I

Marco teórico referencial para el empleo de la minería de datos en las bibliotecas

Este capítulo presenta las herramientas, metodologías, técnicas, procesos, que son útiles para la creación de la metodología que se propone, para el estudio de la información bibliográfica.

Tanto en el ámbito internacional como nacional, existen muchas publicaciones sobre metodologías, software y técnicas usadas para el descubrimiento de conocimiento, con el empleo de la minería de datos.

Estas publicaciones abarcan diversos temas, desde las redes neuronales artificiales (RNA), máquinas que aprenden utilizando estas RNA, inteligencia artificial (IA), recuperación de información (RI), clusterización (C), data mining (DM), text mining (TM), web data mining (WDM), data warehouse, online analytical processing (OLAP), online public access catalog (OPAC) entre otros y todas mencionan la obtención de conocimiento oculto a través de herramientas digitales, que se pueden adquirir por medio de Internet, algunas de forma pagadas y otras exentas de pago. Son software que utilizan algoritmos en sus ejecuciones para extraer patrones, reglas, asociaciones, correlaciones e incluso interesantes excepciones potencialmente útiles, que pueden ser desconocidas y estar ocultas en bases de datos, bancos o repositorios (Arcos et al., 2019), (Nieto, 2021).

La revisión de tres herramientas conocidas para el manejo de grandes bases de datos como son la data warehousing, el OLAP y el OPAC permitió conocer que:

- *data warehousing*, están consideradas como un conjunto de datos integrados, históricos, variantes en el tiempo y unidos alrededor de un tema específico, que es usado por la gerencia para la toma de decisiones, mantiene una arquitectura multidimensional, porque relacionan muchas bases de datos que almacenan en su interior, funcionando de forma transparente para el sistema operativo donde se encuentran (Jeré, 2018).

- *OLAP*, definidos como una solución utilizada en el campo de la llamada Inteligencia empresarial, cuyo objetivo es agilizar la consulta de grandes cantidades de datos, y para ello utiliza estructuras multidimensionales, al igual que la data warehousing, OLAP pone una vista interactiva de los datos empresariales para la toma de decisiones. (Peña, 2015).

- *OPAC*, son los conocidos catálogos automatizados de acceso público en línea, que utilizan las bibliotecas para consultar grandes colecciones de estar disponibles, son el núcleo central de los sistemas de gestión bibliotecario. Estos catálogos tienen la dificultad que el usuario debe saber acotar la consulta que le hace al OPAC, porque si la consulta es demasiado general, recibe una lista extensa de fichas bibliográficas, como sucede en los buscadores actuales (De Volder, 2005).

En resumen, la data warehousing, el OLAP y el OPAC, son tres herramientas muy poderosas para el tratamiento de la información, cada una con sus características que las hacen ser muy diferentes entre ellas. Son herramientas que permiten hacer consultas relativamente estandarizadas, actualizaciones de datos por lotes, generación de cubos de información (o bases de datos multidimensionales), consultas complejas que pueden tomar muchas horas de ejecución, pero ninguno de estos resultados son comparables con los resultados de la minería de datos, ya que la minería de datos procesa cualquier tipo de información, extrayéndole patrones que no forman parte de ningún dato en particular, la minería de datos logra mostrar el poder descriptivo y predictivo de sus resultados (Ibarra, 2021).

1 Antecedentes de enfoques, herramientas y técnicas empleadas para el análisis de la información en las bibliotecas

Algunos años antes que el concepto de la minería de datos se extendiera, autores que trabajaban en el ámbito de las bibliotecas, se habían dado a la tarea de aplicar nuevos enfoques de análisis a la información, con el apoyo de algunas herramientas (Candás, 2006), entre estos autores y sus logros alcanzados figuran algunos como:

- Groszs and Gross (1927) hizo un recuento de las citas para evaluar el uso de la literatura (Urbizagastegui et al., 1998).

- Baughman (1974) analizó 11.130 citas procedentes de 446 revistas, concluyendo que 3.521 citas procedían de 612 revistas diferentes. Aplicando la ley de Bradford[1] a esta literatura, estableció un núcleo de 10 revistas para el campo de la sociología (Urbizagastegui et al., 1998).

- Pao (1975) aplicó la ley de Bradford a la literatura citada por 27 reseñadores en la sección “Bibliography of Medical Reviews” del Index Medicus de 1967 a 1970, referido a drogas terapéuticas, usadas para combatir la arritmia cardiaca. Ello le permitió la identificación de un núcleo de 14 artículos básicos, 6 artículos de revisión y 3 textos clásicos en el campo de la cardiología y farmacología (Urbizagastegui et al., 1998).

- Pan (1978) encontró que la frecuencia con que una revista es citada es tan confiable para predecir el uso de esas revistas en la biblioteca, como la opinión de los bibliotecarios familiarizados con esas revistas y sus usuarios (Urbizagastegui et al., 1998).

- Nutter (1987) exploró las fuentes de datos de la biblioteca para apoyar la toma de decisiones, pero lamentó que "la capacidad de recopilar, organizar y manipular los datos era muy superior a la capacidad de interpretar y aplicar de ellos" (Aguilar, 2016).

- White (1987***)*** analizó 1.114 citas tomadas de la base de datos CONSULTANT e identificó 9 revistas responsables del 80% de las citas sobre enfermedades en pequeños animales y 14 revistas responsables del 80% de las citas sobre enfermedades en animales de tamaño mayor (Urbizagastegui et al., 1998).

- Johnston y Weckert (1990) desarrollaron un sistema experto basado en datos para ayudar a seleccionar los materiales de biblioteca (Nicholson, 2003).

1. **Ley de Bradford** es un modelo descrito originalmente por Samuel C. Bradford en el año 1934, para estimar la disminución exponencial de rendimiento (decreciente) y ampliar la búsqueda de las referencias en las revistas científicas. http://www.historyofinformation.com/detail.php?entryid=729

- Vizine-Goetz, Weibel, y Oskins (1990) desarrolló un sistema para la catalogación automatizada basado en títulos de libros (Nicholson, 2003).

- Fayyad y Simoudis (1995) definían minería de datos (aunque predominaba el uso de la expresión *knowledge discovery in databases* como "un campo emergente que combina técnicas de aprendizaje-máquina, reconocimiento de patrones, estadística, bases de datos y visualización para extraer automáticamente conceptos, conceptos interrelacionados, y patrones de interés desde grandes bases de datos (Candás Romero, 2006).

- Mancini (1996) planteó como extraer de los datos, información útil para la toma de decisiones a nivel de dirección de una biblioteca (Nicholson, 2003).

- Peters (1996) planteó como examinar los datos extraídos de la colección, para apoyar en la administración de los fondos de la biblioteca (Nicholson, 2003).

- Banerjeree (1998) describió el proceso de minería de datos y la forma de usarlo, para proporcionar un acceso mejor a la colección (Nicholson, 2003).

- Schulman (1998) utilizó la minería de datos para examinar las tendencias sobre el comportamiento del usuario de la biblioteca (Nicholson, 2003).

- Guenther (2000) discute sobre las fuentes de datos y aplicaciones de bibliominería, pero se enfoca en los problemas con los formatos heterogéneos de datos (Nicholson et al., 2003).

- Liddy (2000) combinó el proceso del lenguaje natural con la minería de datos textual, para descubrir la información oculta, en colecciones de las bibliotecas digitales (Nicholson, 2003).

- Chau (2000) planteó el uso de la explotación de la minería de datos en la Web, para personalizar los servicios de referencia electrónica (Nicholson, 2003).

- Doszkocs (2001) planteó aplicar redes neuronales a los datos de la biblioteca, para descubrir posibles asociaciones entre los documentos, la indexación de términos, los códigos de clasificación y las consultas (Nicholson, 2003).

Un análisis de los enfoques, herramientas y técnicas empleadas por estos investigadores con el estudio de los datos de las bibliotecas, permite resumir que sus resultados han contribuido al mejoramiento de las colecciones de los fondos, la toma de decisiones para mejorar el cuadro de mando, las citas de las publicaciones para conocer su impacto, y en años más recientes descubrir el comportamiento de los grupos de usuarios, el comportamiento del propio personal de las bibliotecas o de ambos (De la Puente, 2010) o también sobre estudios denominados como educational data mining (EDM), el cual trata de predecir el comportamiento futuro de los estudiantes en relación a los efectos del soporte pedagógico (Yu et al., 2021).

1.1 Estudios de las fuentes seleccionadas con minería de datos

La necesidad de aplicar estos análisis de minería en la creciente documentación, que se va generando en todos los sectores donde trabaja el hombre, provocó que se comenzara a manejar el termino de minería de textos alrededor de los años ochenta (Valero, 2017) quedando desde entonces establecido que la minería de datos es el proceso que se aplica cuando la información está estructurada o proveniente de una base de datos, mientras que la minería de textos, es el proceso que se aplica cuando la información a procesar es información no estructurada, es decir proviene de las colecciones de textos o volumen de documentos (Zambrano et al., 2021).

Se puede concluir que la minería de textos y la recuperación de información son procesos totalmente diferentes, la “recuperación de información” consiste, en la recuperación automática de documentos relevantes mediante un buscador que hace indexaciones de textos, clasificación, categorización, mientras que la “minería de textos”, es el proceso de extraer información que no está contenida en ningún texto en específico, sino que es la información global y desconocida sobre el comportamiento que tienen todos los registros, textos o documentos de la colección que se esté analizando (Chang et al., 2018).

Para el término de d*escubrimiento de conocimiento en base de datos* o *knowledge discovery in data bases* (KDD) como también se menciona en la literatura (Wirth et al., 2005)**,** se encuentran varias definiciones dada por diferentes autores como:

- Según *Michalski* (1986) el *descubrimiento de conocimiento* es un tipo de inducción de conocimiento, no supervisado, que implica dos procesos: (Gómez, 2014; Michalski, 1986).

- búsqueda de regularidades interesantes entre los datos de partida,
- formulación de *leyes* que las describan.

- Según *Walker* (19*87)*, el *descubrimiento de conocimiento* implica observar, recoger datos, formar hipótesis para explicar las observaciones, diseñar experimentos, comprobar la corrección de las hipótesis, comparar nuestros hallazgos con los de otros investigadores y repetir el ciclo (Gómez, 2014; Walker, 1987).

- Según Frawley (1992) el *descubrimiento de conocimiento* es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de un conjunto de datos (Frawley et al., 1992; Gómez, 2014).

- Según Fayyad (1996) el *descubrimiento de conocimiento* es el conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objetivo de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos (Fayyad et al., 1996; Gómez, 2014).

La definición de Frawley en 1992 ha resultado ser la más difundida y utilizada, sobre el descubrimiento de conocimiento aplicado a la información y también la que se asume como definición válida para esta investigación. Para dar respuesta a los objetivos propuestos en esta investigación sólo se consideran patrones potencialmente útiles a aquellos patrones, que de su análisis se puede extraer conocimiento, es información que no está guardada en ningún campo de la base de datos y que solo a través del análisis de esos patrones se puede conocer su comportamiento.

En la revisión bibliográfica realizada se encontraron ocho trabajos que se acercan a los objetivos de esta investigación asociados a los términos de bibliominería, minería de datos y minería de textos, cuatro son trabajos foráneos y cuatro trabajos nacionales.

El análisis de los cuatro trabajos foráneos se resume de la siguiente manera:

- Hernán Merlino, en su artículo **'METODOLOGÍA DE TRANSFORMACION DE DATOS PARA SU EXPLOTACIÓN'**, se refiere a *"un método de transformación de datos orientado a la explotación de información, y se detallan las características necesarias que debe poseer el entorno de trabajo, para la automatización del mismo"* (Merlino, 2004).

Este trabajo presenta el proceso de selección y transformación de los datos y de las condiciones externas a la información que deben existir, para someterlos a un proceso de minería exitoso. No crea, ni valida metodología alguna en particular a pesar de su título, es un trabajo totalmente útil y teórico para esta investigación, ya que hace énfasis en la forma de trabajo con los datos.

- Scott Nicholson, creador del término bibliomining; y del artículo **'THE BASIS FOR BIBLIOMINING: FRAMEWORKS FOR BRINGING TOGETHER USAGE-BASED DATA MINING AND BIBLIOMETRICS THROUGH DATA WAREHOUSING IN DIGITAL LIBRARY SERVICES'**, *analiza la integración de la minería de datos en los servicios de la biblioteca digital. En primer lugar, bibliominería, o la combinación de la bibliometría y técnicas de minería de datos para entender los servicios de biblioteca, y se define el concepto explorado. En segundo lugar, los marcos conceptuales para bibliominería desde el punto de vista de la biblioteca de toma de decisiones y el investigador de la biblioteca son presentados y comparados.* (Nicholson, 2006).

Este trabajo es una guía sobre el tratamiento de la información e implementación de los procesos de minería de datos en un OPAC. Explica sobre la implementación de una data warehouse, y como esta herramienta necesita de una estructura que reúna información de tres tipos, una data sobre las publicaciones, otra data sobre los usuarios y otra data sobre los servicios. Hace mención del proyecto digital reference electronic warehouse (DREW) y de los consorcios entre bibliotecas y menciona formas de trabajos entre ellas, aunque no propone ninguna metodología para la implantación de estos procesos.

- Jorge Candás Romero en su trabajo llamado **'MINERÍA DE DATOS EN BIBLIOTECAS: BIBLIOMINERÍA'**, *"presenta una introducción teórica a la*

aplicación de la minería de datos en bibliotecas, denominada bibliominería. Asimismo, se presentan algunas de las posibles aplicaciones prácticas y cómo éstas sirven de apoyo a la llamada Biblioteca 2.0 y a la creación y gestión de servicios más y mejor orientados al usuario, basados en nuevas tecnologías. Finalmente se analiza el problema de la privacidad en la aplicación de la bibliominería" (Candás, 2006).

Es una introducción teórica a la minería de datos para profundizar después en su aplicación en bibliotecas, presenta teorías y conceptos de algunas aplicaciones prácticas utilizadas hasta el momento, en proyectos de este tipo y en relación a las nuevas definiciones de Web 2.0, aunque no propone ninguna metodología, ni tampoco ejecuta ningún proceso de minería de datos.

- Ricardo Herrera Varela, en su tesis de doctorado, titulada **'BIBLIOMINING: MINERÍA DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS APLICADOS AL ÁMBITO BIBLIOTECARIO'**, *"se estudian las fases del proceso de descubrimiento del conocimiento en bases de datos (KDD), que incluyen datos de tipo complejo y que tiene a la minería de datos, como elemento esencial para apoyar este proceso mediante el análisis de datos. Se analizan el concepto y desarrollo de las fases del proceso de bibliomining, como algunas aplicaciones y beneficios, que pude significar esta disciplina para el ámbito bibliotecario. Finalmente se intenta esbozar algunas consideraciones a tener en cuenta antes de desarrollar un proyecto de bibliomining" (Herrera Varela, 2006)*.

Este trabajo es muy útil porque *describe los pasos del proceso de extracción de conocimiento en bases de datos y como se pueden utilizar la explotación minera de los datos en bibliotecas, para entender en primer lugar, los patrones de comportamiento entre usuarios y el personal de la biblioteca, y en una segunda instancia comprender los patrones de uso de los recursos de información en la institución*, menciona la metodología CRISP-DM, pero no desarrolla ningún proceso.

Haciendo resumen de estas cuatro publicaciones seleccionados se puede decir que:

- El primer artículo del autor Hernán Merlino, se selecciona porque la teoría que expone coincide con la fase de selección, limpieza y transformación de los datos de nuestra investigación. Este artículo está dedicado a la fase que ocupa más tiempo, esfuerzo y que además puede poner en riesgo, la confiabilidad de los

patrones que se quiere encontrar para mejorar los productos y servicios bibliotecarios.

- El segundo artículo del autor Scott Nicholson, fue seleccionado porque plantea de forma bastante completa, sobre los aspectos que se deben considerar para ejecutar un proceso de minería de datos en el ámbito de las bibliotecas y como el mismo plantea, define los marcos conceptuales para bibliominería desde el punto de vista de la biblioteca.
- El tercer artículo del autor Jorge Candás Romero hace un breve resumen de algunas de las investigaciones realizadas para mejorar los productos y servicios bibliotecarios y nos resume el trabajo de algunos de los investigadores y de las nuevas técnicas en las bibliotecas versión 2.
- El cuarto artículo del autor Ricardo Herrera Varela plantea teóricamente, sobre los pasos para desarrollar un proceso de minería de datos, la metodología CRISP-DM, sus creadores y el entorno donde con mayor éxito ha sido aplicada, también hace un resumen en tabla asociando las tareas de modelación con las técnicas de minería de datos que se usan frecuentemente.

Cada uno de estos trabajos aportaron procedimientos, herramientas, proyectos y metodologías reconocidas, que sirvieron para conformar el mapa del conocimiento foráneo asociado a los objetivos de esta investigación, de igual manera se procedió para conocer en el ámbito nacional sobre lo que se ha trabajado, en el tema de la minería de datos aplicada al entorno bibliotecario.

En cuanto al ámbito nacional, en Cuba existen dos centros nacionales, dedicados al estudio e investigación de los procesos de minería de datos. Uno es el Centro de Aplicaciones de Tecnologías de Avanzada' (CENATAV, 2023) y el otro es el 'Centro de Estudios de Reconocimiento de Patrones y Minería de Datos, (CERPAMID, 2023).

En el caso del CENATAV, tiene un área de investigación dedicada al desarrollo de investigaciones teóricas y aplicadas en el área del reconocimiento de patrones y la minería de datos**,** desde el año 2003, y ambas instituciones trabajan de conjunto con la Universidad de Oriente, que a su vez trabajó unida a la Universidad de Jaume I, Castellón en España y de aquí se escogieron trabajos nacionales.

Entre los trabajos cubanos también se escogieron cuatro investigaciones:

- De los autores Damaris Pascual, Filiberto Pla, J. Salvador Sánchez, se encuentra el trabajo realizado de conjunto con la Universidad Jaume I, (España): **"HIERARCHICAL-BASED CLUSTERING USING LOCAL DENSITY INFORMATION FOR OVERLAPPING DISTRIBUTIONS"**, que plantea, "*las técnicas de clusterización son ampliamente usadas en muchos campos de aplicación como el análisis de imágenes, la minería de datos y el descubrimiento de conocimiento entre otros. En este trabajo, se presenta un nuevo algoritmo de clusterización para encontrar clústeres de diferentes tamaños, formas y densidades capaces de tratar con la superposición de distribuciones de clúster y el ruido de fondo. El algoritmo está dividido en dos etapas, en la primera etapa, la densidad local se calcula en cada punto de los datos. Esta densidad local es usada para inicializar la clusterización agrupando los objetos alrededor del objeto de máxima densidad local (punto central). En el segundo paso, un enfoque jerárquico es usado mediante la fusión de los clústeres de acuerdo a la distancia del clúster introducido, también basada en la información de la densidad local. Resultados experimentales en bases de datos sintéticos y reales muestran la validez del método propuesto"* (Pascual et al., 2006).

Este artículo trata sobre la técnica de clusterización y las conclusiones que expone sobre enfoque jerárquico, clústeres y densidades, sirvieron para entender mejor los resultados de la clusterización, que se logran en nuestra investigación.

- El artículo titulado **"UNA PROPUESTA BASADA EN LA ESTIMACIÓN DE LAS PROBABILIDADES PARA LA EDICIÓN UTILIZANDO EL CLASIFICADOR K-NN"**, de los autores F. Vázquez; F. Pla, J. S. Sánchez, expone que "*los clasificadores supervisados basan su aprendizaje en un conjunto de datos denominado conjunto de entrenamiento, mediante el cual, se proporciona al clasificador una serie de casos o situaciones con las que puede encontrarse al requerirse una predicción o clasificación de un nuevo objeto. La idea central del presente artículo es utilizar como regla de clasificación aquella que utiliza la estimación de la probabilidad de pertenencia a la clase de cada uno de los k vecinos más cercanos. A partir de este esquema de clasificación se implementa la variante repetitiva del algoritmo de Edición de Wilson utilizando la regla de clasificación de Centroide más Próximo (Wilsoncn), así como también, la de los*

algoritmos de Wilson con probabilidades y Wilson con probabilidades y umbral. Todos los algoritmos propuestos se comparan con algunas de las más populares técnicas de edición reportadas en la literatura; como son Edición por Partición, Multiedit y algoritmo de Wilson" (Vázquez et al., 2007).

Este artículo presenta un estudio sobre clasificación usando *los k vecinos más cercanos*, y a todos los algoritmos propuestos los comparan con algunas de las más populares técnicas de edición reportadas en la literatura; como son Edición por Partición, Multiedit y algoritmo de Wilson. Con sus conclusiones, facilita la comprensión de los resultados de nuestra investigación.

- De los autores Guillermo Matos, Ricardo Chalmeta y Oscar Coltell se escogió el trabajo "**METODOLOGÍA PARA LA EXTRACCIÓN DEL CONOCIMIENTO EMPRESARIAL A PARTIR DE LOS DATOS"**, "*en este trabajo se presenta una metodología formal para la extracción del conocimiento a partir de los datos que dispone una empresa. Los datos que residen en las bases de datos corporativas pueden ser una de las fuentes de conocimiento más importantes que hay en las empresas por lo que su manejo eficiente es de especial importancia. La metodología propuesta se ha denominado KM-IRIS y consta de cinco fases. En cada fase se proponen los objetivos a cumplir y las técnicas y herramientas que se pueden aplicar. Este planteamiento facilita la aplicación inmediata de la metodología siguiendo simplemente las pautas expuestas. Con el propósito de calibrar y validar la metodología, se ha aplicado el método desarrollado a varios casos típicos, como pueden son la extracción del conocimiento a partir de las personas, los documentos o los datos"* (Matos et al., 2006).

Este artículo presenta una metodología conocida por el acrónimo de *KM-IRIS* que utiliza datos empresariales, además que para lograr los resultados de la fase Extraer usan una data warehouse federado con un repositorio de conocimiento y para desarrollar la fase *Procesar,* presenta una tabla que tiene las herramientas y plataforma que los autores proponen. Toda esta propuesta se toma como experiencia para esta investigación.

- De los autores Dr.C. Juan Pedro Febles Rodríguez y Abel González Pérez, el trabajo **'APLICACIÓN DE LA MINERÍA DE DATOS EN LA BIOINFORMÁTICA'**, plantea que "*en los próximos años ocurrirá un avance espectacular de las ciencias*

biomédicas como resultado del proyecto Genoma Humano. Las nuevas tecnologías, basadas en la genética molecular y la informática, son claves para este desarrollo, pues ellas suministran potentes instrumentos para la obtención y el análisis de la información genética. La aparición de nuevas tecnologías ha posibilitado el desarrollo de la genómica, al facilitar el estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades, todo lo cual influye en el diagnóstico clínico, la investigación de nuevos fármacos, la epidemiología y la informática médica. En los últimos años, la minería de datos ha experimentado un auge como soporte para las filosofías de la gestión de la información y el conocimiento, así como para el descubrimiento del significado que poseen los datos almacenados en grandes bancos. Esta permite explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones; además de facilitar la extracción de la información existente en los textos, así como crear sistemas inteligentes capaces de entenderlos, a esto se denomina comúnmente como minería de textos (textmining). Se describen sintéticamente los componentes básicos de la minería de datos y su aplicación en una emergente y trascendental actividad científica: la bioinformática" (Febles Rodríguez et al., 2002).

En este artículo los autores identifican *tres componentes básicas de los métodos de la minería de datos*, y la definen de la siguiente manera:

- ***Lenguaje de representación del modelo**:* comprende las suposiciones y restricciones utilizadas en la representación empleada.
- ***Evaluación del modelo:*** incluye el uso de técnicas de validación cruzada para la predictividad y aplicación de principios como el de máxima verosimilitud o el de descripción mínima para evaluar la calidad descriptiva del modelo.
- ***Método de búsqueda:*** puede dividirse en búsqueda de parámetros y búsqueda del modelo, determinan los criterios que se siguen para encontrar los modelos.

Este trabajo es seleccionado por definir las componentes básicas de los métodos de la minería de datos, lo cual se tendrá en cuenta durante la confección y validación de la metodología que se está desarrollando.

En resumen, se seleccionaron estos cuatro artículos nacionales porque dos artículos están enfocados al estudio de los algoritmos de la minería y al tipo de resultado que se obtiene de ellos, mientras que los otros dos están dedicados a simplificar el

proceso de descubrimiento de conocimiento, que es uno de los objetivos de esta investigación.

1.2 Propuesta para desarrollar un proyecto de descubrimiento de conocimiento

Existen varios procedimientos para desarrollar un proyecto de descubrimiento de conocimiento, pero la aceptada para esta investigación es la planteada por Cabena y otros investigadores (Cabena et al., 1998), porque solo consta de 5 procesos que son muy sencillos de implementar, aun por personal no especializado y estos procesos se ordenan de la siguiente manera:

1. **Determinación de los Objetivos.**
2. **Preparación de datos.**
   a. Selección: Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.
   b. Pre procesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
   c. Transformación de datos: conversión de datos en un modelo analítico.
3. **Minería de datos.**
   a. Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos.
4. **Análisis de Resultados.**
   a. Interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.
5. **Asimilación del conocimiento.**
   a. Aplicación del conocimiento descubierto.

Estos cinco procesos presentan etapas que recogen de una forma muy desglosada y sencilla, todo el trabajo que hay que realizar para encontrar los patrones ocultos, que se encuentra en la información y que sirven para demostrar la validez de la metodología que se presenta.

1.3 Tareas y técnicas asociadas a la minería de datos

Todas las herramientas digitales de código abierto empleadas para realizar la fase de minería de datos, contienen los algoritmos matemáticos para desarrollar todas las tareas y técnicas propias de un proceso de minería de datos. Solo se debe tener en

cuenta que el correcto funcionamiento de estos algoritmos, depende del tipo de información que se esté procesando (Romero, 2019).

Es necesario recordar que el pre procesamiento de los datos en el punto b, incluye tipo de información, la estandarización de la misma, la toma de decisiones sobre los campos vacíos, los registros repetidos o incompletos, los ruidos, la selección de las variables representativas, entre otros, todo antes de importar la información al software. Esta es la única forma para lograr que cualquiera de los algoritmos de minería de datos funcione correctamente y que se obtengan los patrones de comportamiento de las variables que fueron escogidas, de los cuales tras previo análisis se infiere el conocimiento necesario, para dar respuesta a los objetivos que conllevaron a realizar este proceso (González López, 2021).

Las tareas y técnicas empleadas en los procesos de minería de datos en esta investigación se resumen en la Tabla1.

No.	**Tareas de Modelación**	**Algunas Técnicas asociadas a la Minería de Datos**
1	Clasificación	Métodos de inducción de reglas, Árboles de Decisión, K vecinos más cercanos.
2	Predicción	Análisis de regresión, Árboles de regresión, Redes Neuronales, K vecinos más cercanos.
3	Análisis de Dependencia	Análisis de Correlación, Análisis de Regresión, Redes Bayesianas.
4	Segmentación y Agrupación	Técnicas de Agrupación, redes neuronales, Técnicas de visualización.

Tabla 1. Tareas de modelación y técnicas de minería de datos.
Fuente: Herrera Varela, R. Diciembre2006. **"*Bibliomining: minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario*"**

Como se observa, son cuatro las *'Tareas de Modelación'* encargadas de agrupar las técnicas de minería de la información más usadas. Estas técnicas responden a algoritmos matemáticos específicos, por lo que se agrupan en dos grandes categorías como son, los *algoritmos supervisados o predictivos* y los algoritmos *no supervisados o de descubrimiento del conocimiento* (Sancho, 2020b).

Caracterizando brevemente cada una de las tareas de modelación y las técnicas de minería de datos, se tiene que:

1. La **Clasificación**, está considerada como tarea de modelación, es una de las formas de análisis más antiguo usado por el hombre para agrupar elementos que comparten propiedades comunes (Britez, 2021). El conocimiento de las propiedades de estos grupos o clústeres, permite una descripción sintética, de un conjunto de datos multidimensional complejo. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo, por la de un representante característico del mismo.
Dentro de las técnicas que se agrupan bajo el concepto de clasificación o clustering se encuentra:

Métodos de inducción de reglas (Sancho, 2021), (IBM, 2021b)
El aprendizaje inductivo es un caso particular entre las técnicas de aprendizaje a partir de ejemplos, siendo su cometido el inducir reglas a partir de los datos disponibles, para lo cual procederá a clasificar en la clase correspondiente diferentes objetos, basándose en el valor de las características o atributos que los definen (Roman, 2019).

Como resultado de una clasificación basada en reglas (Hipp et al., 2002), se encuentran los que representa la relación existente entre la conclusión-decisión y sus atributos. Entre estos métodos está el árbol de decisión que clasifica correctamente los ejemplos dados y minimiza el número de atributos requeridos para alcanzar la conclusión-decisión, siendo esta la explicación de por qué ciertos atributos no aparecen en el árbol.
También están los Dendrogramas, que fundamentalmente usan argumentos de naturaleza textual y permiten una forma de organización jerárquica de todo el dominio(IBM, 2023a).

Árbol de Decisión (Salazar et al., 2021)
Los árboles de decisión (también conocidos como árboles de clasificación y regresión) son los pilares tradicionales de minería de datos y uno de los clásicos algoritmos de aprendizaje de máquina. Desde su desarrollo en la década de 1980, lo más utilizado por la minería de datos ha sido el aprendizaje de maquina basados en el constructor de modelos. La atracción reside en la simplicidad del modelo resultante, donde un árbol de decisión es bastante fácil de ver, entender y explicar (Martínez Heras, 2021).

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, los nodos del árbol están etiquetados con nombres de atributos, las ramas con los posibles valores del atributo, y las hojas con las diferentes clases individuales (Flores Rodrıguez, 2021). Este método tiene un carácter no paramétrico, y no precisa de hipótesis preestablecidas sobre las variables de partida (Migriño et al., 2021). Por eso los árboles de decisión o clasificación son uno de los métodos más claros y difundidos por estar fuertemente basada en los valores de sus atributos, y no en las relaciones establecidas entre estos.

K vecinos más cercanos (Merkle, 2021)

El método K vecinos más cercanos, es un método de clasificación supervisada, es el método más básico de discriminación no paramétrico, ya que no se hace ninguna suposición distribucional. El algoritmo *k-nn* es usado como método de clasificación de elementos, está basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos, es decir aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos, sirve para estimar la función de densidad (Mantilla et al., 2021).

Este método de clasificación no paramétrico, estima el valor de la función de densidad de probabilidad, o directamente la probabilidad, a partir de la información proporcionada por el conjunto de prototipos, sus reglas de clasificación por vecindad, están basadas en la búsqueda de un conjunto de los k prototipos más cercanos, al patrón a clasificar (Sancho, 2020b).

2. La **Predicción**, como su nombre lo indica, es una declaración precisa de lo que ocurrirá en determinadas condiciones especificadas. La Predicción es muy similar a la clasificación, la única diferencia es que en la predicción el atributo objetivo (la clase), no es un atributo cualitativo discreto sino uno continuo, porque el objetivo de la predicción consiste en encontrar el valor numérico del atributo objetivo para objetos no vistos. En la literatura, este tipo de problema es a veces llamado regresión. Si la predicción trata con datos de serie tiempo, entonces a menudo lo llaman pronóstico (Vargas, 2020).

Dentro de las técnicas de minería de datos que se agrupan bajo la Predicción se encuentran:

Análisis de regresión (Díaz, 2021)

El análisis de regresión es uno de los métodos supervisados más utilizados en la Estadística Multivariante, el estudio de la regresión nos permite averiguar, hasta qué punto una variable puede ser prevista conociendo otra, es decir, se utiliza para intentar predecir el comportamiento de ciertas variables a partir de otras, así *el análisis de regresión sirve tanto para explorar datos, como para confirmar teorías*. Existen varios tipos de regresión, por ejemplo:

- *Regresión lineal simple*-(Sólo se maneja una variable independiente, por lo que sólo cuenta con dos parámetros).
- *Regresión lineal múltiple*-(Maneja varias variables independientes, cuenta con varios parámetros).
- *Regresión logística binaria*-(Modelo de regresión para variables dependientes o de respuesta binomialmente distribuidas). Los modelos de regresión logística permiten estudiar si una variable binomial depende, o no, de otra u otras variables (no necesariamente binomiales) (Martínez et al., 2020).

Árbol de regresión (Maydana, 2021)

Los árboles de clasificación utilizan una regresión lineal, para predecir los valores de las clases, mientras que los árboles de regresión, se usan para predecir valores promedio de los valores en las hojas. Los valores que pueden tomar las entradas y las salidas, pueden ser valores discretos o continuos y se utilizan más los valores discretos por simplicidad. Cuando se utilizan valores discretos en las funciones de una aplicación, se denomina clasificación y cuando se utilizan los valores continuos se denomina regresión (Calvo Pérez, 2021). El interés del árbol es la ordenación de la población por grupos y estos grupos se pueden entender y localizar en la base de datos.

Redes Neuronales (IBM, 2023b)

Las redes de neuronas artificiales (denominadas habitualmente como RNA o en inglés como: "ANN", son una simulación de las propiedades observadas en los sistemas

neuronales biológicos y representados a través de modelos matemáticos (Sancho, 2019a). En lo que se refiere a inteligencia artificial las redes neuronales, son una representación de aprendizaje y procesamiento automático. Las redes neuronales pueden utilizarse en diferentes áreas como, explotación de bases de datos, reconocimiento de caracteres escritos, análisis de tendencias y patrones.

Entre las llamadas redes neuronales están los mapas auto-organizados o Self Organizing Map (SOM), también conocidos como Redes de Kohonen, son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de dos dimensiones normalmente, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella (Ruiz Varona et al., 2020).

A lo largo del entrenamiento de la red, los vectores de datos son introducidos en cada neurona, y se comparan con el *vector de peso* característico de cada neurona. La neurona que presenta menor diferencia, entre su vector de peso y el vector de datos, es la neurona ganadora (o *BMU*, Best Matching Unit) y ella y sus vecinas, verán modificados sus vectores de pesos (Espinoza Hoyos, 2020).

3. El **Análisis de Dependencias** consiste en encontrar un modelo que describa dependencias significativas (o asociaciones) entre artículos de datos o acontecimientos, se dice que una clasificación extensa, divide las dependencias en dos tipos, dependencias de control y dependencias de datos.

Las dependencias pueden ser estrictas o probabilísticas y aunque las dependencias pueden ser usadas para el modelado predictivo, son más utilizadas en la clasificación por su comprensión, entre las técnicas de minería de datos asociadas a esta tarea encontramos:

Análisis de Correlación (Roy et al., 2019)
El Coeficiente de Correlación, es la medida de la intensidad de la relación lineal entre dos variables, puede tomar valores desde menos uno hasta uno, indicando que mientras más cercano a uno sea el valor del coeficiente de correlación, en cualquier dirección (positivo o negativo), más fuerte será la asociación lineal entre estas dos variables.

Comparando la Técnica de Análisis de Correlación y Análisis de Regresión, ambas se consideran como una metodología estadística para predecir hechos y eventos. Con respecto al Análisis de Regresión lo que se hace es evaluar la contribución de una o más variables con respecto de otra, es decir éste análisis permite evaluar que tan bien una o más variables (independientes) ayudan a explicar a otra (dependiente), (Gea et al., 2014).

Redes Bayesianas (Mateos, 2021)
Una red bayesiana o red de creencia, es un modelo probabilístico multivariado, que relaciona un conjunto de variables aleatorias, mediante un grafo dirigido que indica explícitamente influencia causal. Gracias a su motor de actualización de probabilidades, las redes bayesianas son una herramienta extremadamente útil, en la estimación de probabilidades ante nuevas evidencias, una red bayesiana es un tipo de red causal. Un híbrido de red bayesiana y la Teoría de la Utilidad es un diagrama de influencia.

Existen algoritmos que realizan inferencias y aprendizaje basados en redes bayesianas. El obtener una red Bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico (Angulo, 2020).

4. La **Segmentación y Agrupación,** la misma apunta a la separación de los datos en subgrupos o clases significativas e interesantes. Todos los miembros de un subgrupo comparten características comunes.
La segmentación puede ser realizada de forma manual o automática, es un paso hacia la solución de otros tipos de problemas (Sánchez, 2021). Esta tarea tiene 3 técnicas asociadas, las Redes neuronales (explicadas anteriormente), las Técnicas de Agrupación y las Técnicas de Visualización:

Técnicas de Agrupación (Benitez, 2005)
Un algoritmo de agrupamiento (*clustering)*, es un procedimiento de agrupación de una serie de vectores, de acuerdo con un criterio de cercanía. Esta cercanía se define en términos de una determinada función de distancia, aunque existen otras más robustas o que permiten extenderla a variables discretas.

Generalmente, los vectores de un mismo grupo (o *clúster*) comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. Esta descripción sintética se consigue, sustituyendo la descripción de todos los elementos de un grupo, por la de un representante característico del mismo. Existen diversas implementaciones de algoritmos concretos para desarrollar estas técnicas de agrupamiento, por ejemplo, el de las k-medias o de particiones, que es uno de los algoritmos más antiguos y de uso extendido, aunque algunos plantean que posee falta de robustez (Herrero, 2023).

Técnicas de Visualización (Rouse, 2021)
Las técnicas de visualización son consideradas como las técnicas más potentes, para identificar patrones ocultos en los datos. Con estas técnicas se pueden detectar fenómenos que ocurren en los datos, mediante representaciones n-dimensionales de datos sobre pantallas bidimensionales.

Las técnicas de visualización ayudan a presentar y analizar los resultados de los procesos minería de datos, son buenas herramientas que sirven para ubicar patrones en un conjunto de datos, y pueden ser usadas al comienzo de un proceso de minería de datos, para tomar una primera impresión de la calidad del conjunto de datos. Los modelos de visualización pueden ser bidimensionales (2D), tridimensionales (3D) o incluso multidimensionales (Arias, 2021).

Las definiciones anteriores provocan que según los métodos de análisis que emplean estas técnicas, se reagrupen en métodos de análisis *supervisados o no supervisados* (Tabla 2).

Supervisados	**No Supervisados**
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (clustering)
Series temporales	Reglas de asociación
	Patrones secuenciales

Tabla 2. Categoría de las técnicas más usadas en la minería de datos
Fuente: Moreno García, María N. **"Aplicación de técnicas de minería de datos en la Construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software", 2003.**

Los sistemas de clasificación supervisados son aquellos en los que, a partir de un conjunto de ejemplos clasificados (conjunto de entrenamiento), intentamos asignar una clasificación a un segundo conjunto de ejemplos mientras, que los sistemas de clasificación no supervisados son aquellos en los que no disponemos de un conjunto de ejemplos previamente clasificados, sino que únicamente a partir de las propiedades de los ejemplos intentamos dar una agrupación (clasificación, clustering) de los ejemplos según su similaridad (González, 2021).

Existen diversas técnicas de agrupamiento que se dividen en dos grandes categorías:

- Jerárquicas, porque construyen una jerarquía de grupos dividiéndolos iterativamente (Lumbreras, 2020).
- Por particiones, porque el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía (Romero, 2019a).

Los algoritmos *supervisados o predictivos* predicen el valor de un atributo (*etiqueta*), dentro de un conjunto de datos conocidos, es decir a partir del valor de una etiqueta conocida, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos, cuya etiqueta es desconocida.

Cuando una aplicación no tiene el potencial necesario para una solución predictiva, es decir a través de una variable dependiente predecir comportamientos, en ese caso hay que recurrir a los métodos que se conocen como *no supervisados o descubrimiento de conocimiento (KDD)* (Sarker, 2021), para descubrir patrones y tendencias en los datos (no se utilizan datos históricos), en esos casos se considera que no existe una variable dependiente y que todas compiten entre sí bajo algún criterio seleccionado. El descubrimiento de este tipo de información, sirvió para ayudar en el mundo de los negocios y en la investigación (Zhang, 2021).

La caracterización de las Tareas de Modelación y Técnicas de la Minería de datos, permitió seleccionar los algoritmos de la minería que conviene utilizar en esta investigación. Se tuvo en cuenta el tipo de información que se va a procesar, los software a utilizar y la tecnología que se tiene para esto, este previo análisis permitió dar cierto margen de seguridad sobre los resultados que se esperan obtener, además

de facilitar la interpretación de los resultados finales al terminar cada proceso y comprobar si se necesita hacer algunos cambios para comenzar un nuevo proceso.

1.4 Revisión sobre la exploración de los datos y de algunos resultados que se obtienen durante los procesos de Minería de Datos

De la Guía para DATA MINING (Graham Williams, 2010), se seleccionaron algunos términos y definiciones que serán útiles en esta investigación, estas son:

- *Desaparecido (Missing)*

Los valores que faltan dentro de los datos, representan un desafío para la minería de datos y el modelado en general, debido a que provocan patrones falsos, que infieren conocimiento no real al final del proceso de minería de datos. Puede haber muchas razones para los valores que faltan, incluido el hecho de que los datos son difíciles de recoger, no siempre están disponibles, o simplemente que no se registran porque de hecho el valor es cero o no existe. Saber por qué el dato no figura dentro de la base de datos es un elemento muy importante, para conocer cómo tratar con ellos (Emmanuel et al., 2021). Esta información es útil para comprender la estructura de los datos que faltan, y poder hasta llegar a un entendimiento de cómo solucionarlos, porque los algoritmos de Minería de Datos no funcionaran correctamente, sin solucionar este problema previamente (Carpenter et al., 2021).

- Análisis de Agrupamiento (Clustering). (Harmouch, 2021)

El agrupamiento es una de las herramientas básicas utilizadas en la minería de datos, clustering permite agrupar entidades de acuerdo a lo similares que son (Husnain et al., 2021). Esto se hace sobre la base de una medida de la distancia entre las entidades, por lo general el cálculo de las distintas distancias como Euclidean o Spearman (rango basado en métricas), esto se logra haciendo un extracto de los extractos de los campos numéricos, y a continuación el método k-medias (Kumar-Pandey et al., 2021).

Los algoritmos de aprendizaje han sido clasificadas tradicionalmente, en dos grandes categorías: supervisados y no supervisados, dependiendo de si los datos de la etiqueta está disponible o no (Rojas, 2021). Al elegir un tamaño para un clúster sólido necesitamos observar que cuanto mayor sea el número de grupos en relación con el

tamaño de la muestra, entonces se tomará el más pequeño de los grupos, si no hay un tamaño específico de clúster previamente fijado.

- *Gráfica de Distribución (Alcázar Román, 2023)*

Este tipo de gráfico se utiliza como modelo para mostrar el comportamiento de multitud de variables. Los algoritmos que construyen este tipo de gráficas de distribución normal aplican la función para el cálculo del promedio, la desviación estándar a la serie de datos que se está analizando. Es una herramienta que sirve para mostrar el comportamiento de grandes cantidades de información.

- *Histograma (Sotelo, 2020)*

El histograma es una gráfica de la distribución de un conjunto de datos. Un histograma muestra la acumulación o tendencia, la variabilidad o dispersión y la forma de la distribución, sirve para representar variables continuas, aunque también se puede usar para variables discretas.

- *Skewness y Kurtosis (Rigby et al., 2019)*

La *Skewness y la Kurtosis* son medidas estadísticas que se manifiestan a través de gráficas de histogramas. Ambas medidas de dispersión permiten hacer análisis estadísticos, porque caracterizan la ubicación y la variabilidad de un conjunto de datos.

La *Skewness o asimetría* indica la cantidad y dirección de sesgo (alejamiento de la simetría horizontal), es decir de cómo los datos están distribuidos. Una distribución, o conjunto de datos, es simétrica si se ve igual a la izquierda y a la derecha del punto central, su distribución debe estar cerca o ser cero (0), para considerarse una distribución normal (Chen, 2021), mientras que la Kurtosis dice qué tan alta y aguda es el pico de la curva, comparada con la curva estándar de una campana. Los datos que siguen una distribución normal perfectamente tienen un valor de Kurtosis cero(0).

- *Gráfico de dispersión o Scatter Plot* (IBM, 2021a)

Un gráfico de dispersión revela las relaciones o asociación entre dos variables. Estas relaciones se manifiestan por una estructura no aleatoria de la población, es una relación lineal entre las dos variables que indican que un modelo de regresión lineal podría ser apropiado.

También puede ser usada para cruzar más de dos variables, su representación tridimensional (3D) quedaría que en las coordenadas X y Y estarían dos variables independientes y el ploteo sería de una tercera variable dependiente y a la vez agrupada siguiendo el valor de una cuarta variable, (Microsoft, 2021).

- *Mapas Auto-organizados (Self Organizing Maps o SOM)* (Carpio-Martin, 2020)
Los mapas auto-organizados fueron presentados por T. Kohonen en 1982. Este modelo está basado en el tipo de aprendizaje que desarrolla el cerebro humano, que es considerado como un aprendizaje no supervisado competitivo. Así mismo, hace el cerebro durante el proceso de aprendizaje, cuando las neuronas compiten unas con otras con el fin de llevar a cabo una tarea dada y sólo una de las neuronas de salida (o un grupo de vecinas) se activan y al final queda un patrón o grupo de ellos como vencedores, es decir como respuesta y son los que se representan en la red (Gómez-Gil, 2021).

El objetivo de este aprendizaje es categorizar los datos que se introducen en la red, de hecho, las clases o categorías deben ser creadas por la propia red, puesto que se trata de un aprendizaje no supervisado, a través de las correlaciones entre los datos de entrada. Por lo tanto, el SOM es un algoritmo para clasificar observaciones. Su funcionamiento está basado en elegir un gran número de clústeres y colocarlos en forma de una red bidimensional (Sancho, 2021a).

En resumen, de todo lo que ofrece la Guía para DATA MINING, se escogieron y aplicaron en nuestra investigación, solo lo anteriormente mencionado, con el objetivo de cruzar resultados que permitan mejorar los datos y los servicios, todos los resultados son ampliamente explicados en el Capítulo III.

1.5 Metodologías más conocidas, para la implementación de un proceso de Minería de Datos

Dentro de las metodologías existen tres que son muy conocidas y han sido aplicadas en diferentes sectores como el económico, el industrial, la salud y en los negocios, para encontrar patrones útiles (Moine, 2017).
Estas tres metodologías son:

- *CRISP-DM (Cross Industry Standard Process for Data Mining)*

- *SEMMA (Sample, Explore, Modify, Model, Assess)*
- *DMAMC (Definir, Medir, Analizar, Mejorar, Controlar)*

Los orígenes de CRISP-DM *(Cross Industry Standard Process for Data Mining),* se remontan hacia el año 1999, (Wirth et al., 2012), cuando un importante consorcio de empresas europeas proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases), (Fayyad et al., 1996), el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining) (Gallardo Arancibia, 2003) , y que ha sido la más utilizada para la implantación de los procesos de minería de datos según la gráfica publicada en el 2007 por el sitio web KDnuggets.com (KDnuggets, 2007). Esta guía de referencia fue más tarde liberada para su empleo y desarrollo, por parte de la comunidad internacional (Chapman, 2012), y considera seis fases fundamentales para desarrollar un proceso de minería como son: *Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación, Explotación*.

La metodología **CRISP-DM** está dividida en 4 niveles de abstracción, organizados de forma jerárquica y cada uno contiene tareas que van desde el nivel más general hasta los casos más específicos (Gallardo Arancibia, 2003). Esta metodología esta patentizada, con un costo razonable y con un alto impacto en los ambientes académicos e industriales, como metodología no es un procedimiento rígido, se considera una guía que debe ajustarse a cada caso en particular, lo que exige contar con un personal que este bien preparado, para que lo pueda implementar y luego monitorear (E. Fernández et al., 2023).

El Statistical Analysis System Institute Inc., también definió otra metodología con el acrónimo **SEMMA** (Azevedo et al., 2013). El nombre **SEMMA** está conformado con las iniciales de las cinco fases fundamentales que implementa:

- ***S****ample (Muestreo)- (Extrae las muestras para trabajar con ellas),*
- ***E****xplore (Exploración)- (Usa herramientas de visualización y técnicas estadísticas para determinar variables explicativas),*
- ***M****odify (Manipulación)- (Hace formateo de datos),*

- ***M**odel (Modelado)- (Establece relaciones entre variables explicativas y variables objetivos).*
- ***A**sses (Valoración)- (Valorando resultados mediante el análisis de bondad del modelo). Todo este proceso necesita de personal bien preparado para enfrentar estas tareas.*

La metodología **SEMMA** fue creada especialmente para trabajar con las herramientas de minería de datos creadas por y para la propia compañía SAS.

También conocida es otra metodología nombrada **Seis Sigma** o **Six Sigma** (Herrera et al., 2011), sus creadores fueron algunos de los Directores Generales más dotados de Estados Unidos de América, tales como Bob Galvin de Motorola, Larry Bossidy de Allied Signal y Jack Welch de General Electric, estas personas únicamente produjeron guías generales para ser seguidas por los gerentes. **Six Sigma** ofrece técnicas y herramientas para mejorar de manera drástica la calidad de los procesos mientras se reducen los defectos. Es decir, entiéndase como *defecto* cualquier evento en que un producto o servicio no logra cumplir los requisitos del cliente. La meta de Six Sigma es llegar a un máximo de 3,4 *Defectos* por Millón de eventos u Oportunidades (DPMO), sin embargo, muchas empresas han tratado de implementar **Six Sigma** con las cinco fases y los resultados han sido desalentadores, porque en general, los proyectos están ligados a metas de negocios que pueden encontrarse en el Cuadro de Mando (BSC).

En el 1993 de la metodología **Six Sigma** nació la metodología **DMAMC**, con el objetivo de obtener una mayor comprensión sobre los factores críticos que influyen en la excelencia operacional de los países altamente industrializados. **DMAMC** es el acrónimo de *Definir, Medir, Analizar, Mejorar, Controlar*, es en cierta forma, equivalente a PDCA, otro acrónimo como (*Plan, Definir, Chequear, Analizar*), ha sido ampliamente utilizado en resolución de problemas en ambientes (Lara Turrent, 2012).

Existen otras metodologías no tan utilizadas, entre las que se encuentra la metodología **CRITIKAL** (*Client-Server Rule Induction Technology for Industrial Knowledge Acquisition from Large Databases*), fue creada y adoptada por un conjunto de empresas y universidades europeas, y no es de distribución libre. (Al-Attar, 1997).

Esta metodología considera fases muy similares a las de **CRISP-DM** y después de revisar esta y las anteriores, se puede concluir que estas metodologías no cumplen con los propósitos de esta investigación, porque todas están orientadas hacia la industria y los negocios y como **CRITIKAL,** no usa herramientas de distribución libre.

Como el objetivo de esta investigación es el estudio de los datos bibliográficos con minería de datos, se decidió por las herramientas libres que propone el portal líder en KDD de **Gregory Piatetsky** (KDnuggets Polls, 2019). La gráfica sitúa a Rapid Miner V4.6 en segundo lugar, con un 52.7% de utilización y es de obtención libre, mientras que el otro software seleccionado es WinIDAMS13-SP, paquete de análisis estadístico ligero y fácil de usar, que promovió la UNESCO y ahora es posible descargarlo por Softoware.org (Softoware, 2019).

1.6 Política Nacional de Información (PNI) y las bibliotecas

La Política Nacional de Información, aprobada en la Gaceta Oficial de la República de Cuba por el Ministerio de Justicia en Agosto del 2010, reconoce la importancia que tiene la información, cuando es considerada como un recurso económico, productivo y que sirve para el desarrollo de aquellos que la tienen, en forma digitalizada, estandarizada y saben hacer un buen uso de la misma, por lo que orienta específicamente para el caso de las bibliotecas que:

"La Constitución de la República de Cuba en su Artículo 39 establece que el Estado "orienta, fomenta y promueve la educación, la cultura y las ciencias"; y son las bibliotecas, espacios sociales insustituibles para la lectura, la recreación, el aprendizaje continuo, la investigación y el acceso libre a la información que apoyan la formación integral del ciudadano" (MINJUS, 2010).

Estas orientaciones impulsaron a verificar si las bibliotecas cubanas contaban con la información y recursos necesarios para cumplir y las causas que pueden poner en riesgo la investigación y se pudo comprobar que en Cuba, casi todas las bibliotecas cuentan con la implementación de un sistema gestor de información para dar servicio, pero ninguna utiliza las ventajas que proporcionan los patrones extraídos, con las técnicas de minería aplicada a la información bibliográfica, a pesar que con estas técnicas pueden mejorar sus grandes bases de datos e inclusive los resultados de los propios sistemas de recuperación de información.

Para el caso de la Biblioteca de Ciencias y Técnicas se comprueba que también dispone de un sistema gestor, pero como todo sistema gestor de acceso libre no tiene implementado las técnicas de minería de datos, que de poseerlas contribuiría a una vigilancia, inteligencia y prospección más robusta.

Además se verifica que existen varios factores que no ayudan para que las técnicas de minería de datos se incorporen como una tarea cotidiana dentro de la biblioteca:

- No existe una debida integración entre los softwares de minería de datos y la calidad de la información de las bases de datos.
- Se necesitan especialistas que trabajen directamente en el proceso de minería, la interpretación de los resultados y la creación de informes finales que sirvan para hacer propuestas de nuevos productos y servicios.
- No existe la tecnología (hardware) apropiada, con espacio suficiente en disco para el almacenamiento de archivos intermedios, producto de la transformación de los datos, ni con la capacidad y velocidad de memoria (RAM) necesaria, para ejecutar los algoritmos de la minería.
- Algunas de las mejores herramientas dedicadas a ejecutar procesos de minería de datos, son herramientas propietarias y otras aun siendo herramientas libres, no permiten ser descargadas desde Cuba, lo que constituye una barrera tecnológica.

Con el objetivo de presentar las fortalezas y amenazas en esta biblioteca se empleó el diagrama de Ishikawa (causa-efecto) (figura 1).

Diagrama de Ishikawa (causa-efecto)

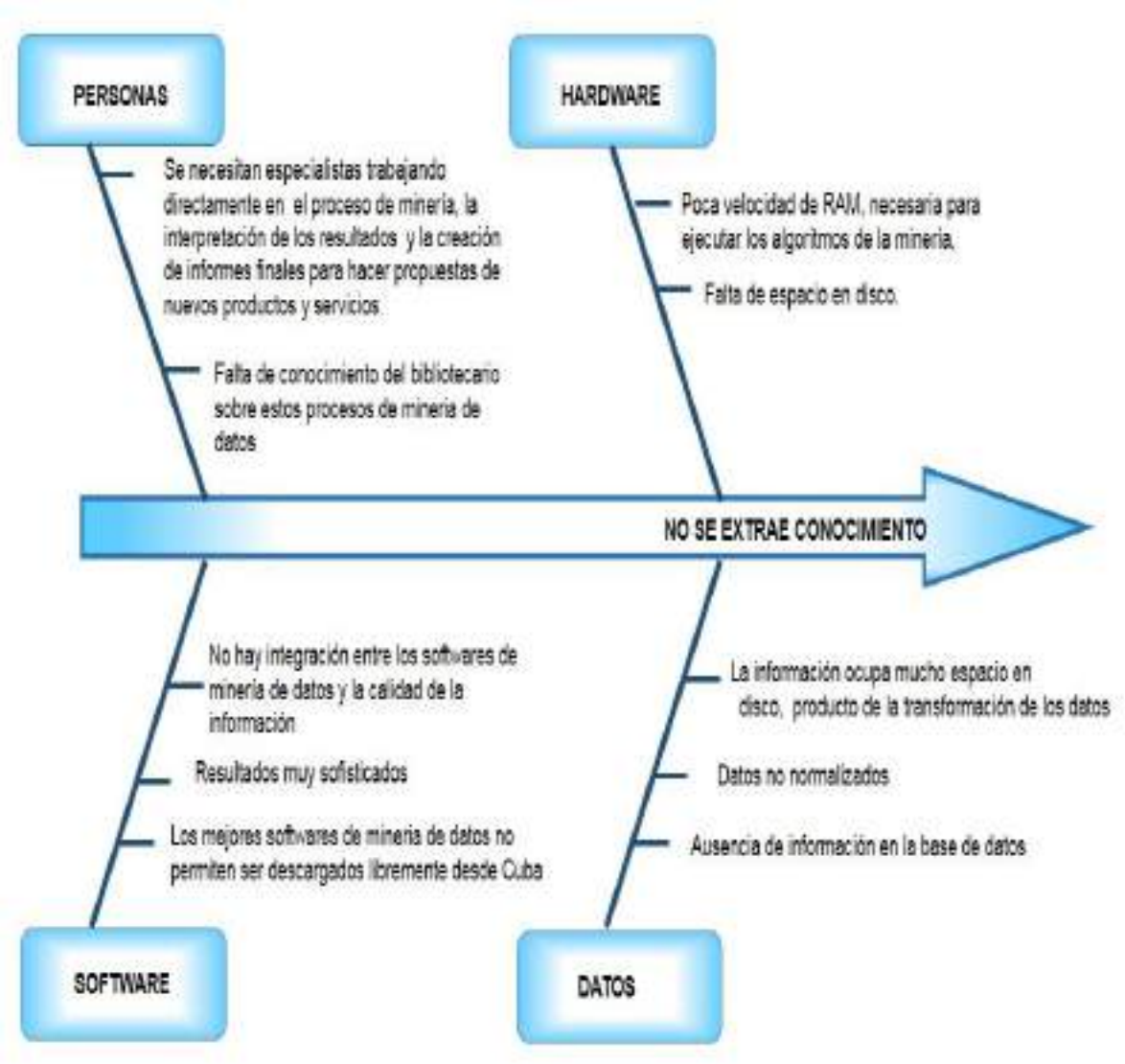


Figura 1 - Diagrama de Ishikawa (causa-efecto)
Fuente: elaboración propia.

Este diagrama de Ishikawa (causa-efecto) fue elaborado a partir del estudio de los 4 factores que tienen más peso dentro de la actividad cotidiana de la biblioteca (Datos, Software, Personas y Hardware) y trabajando de conjunto con los especialista de la biblioteca, se pudo conocer que la ausencia de “Análisis de tendencias y predicciones”, lo convierte en el punto de partida, para que se busquen soluciones con este mismo entorno de trabajo y se pueda brindar mejores productos y servicios de información e inclusive agregar otros nuevos, sustentados con los resultados de la minería de datos. Esto obligó a definir el proceso necesario para desarrollar un proyecto de descubrimiento de conocimiento.

Como resumen de toda la información revisada en la Introducción y el Capítulo I, se creó un mapa conceptual que sintetiza cada uno de estos aspectos vistos (figura 2).

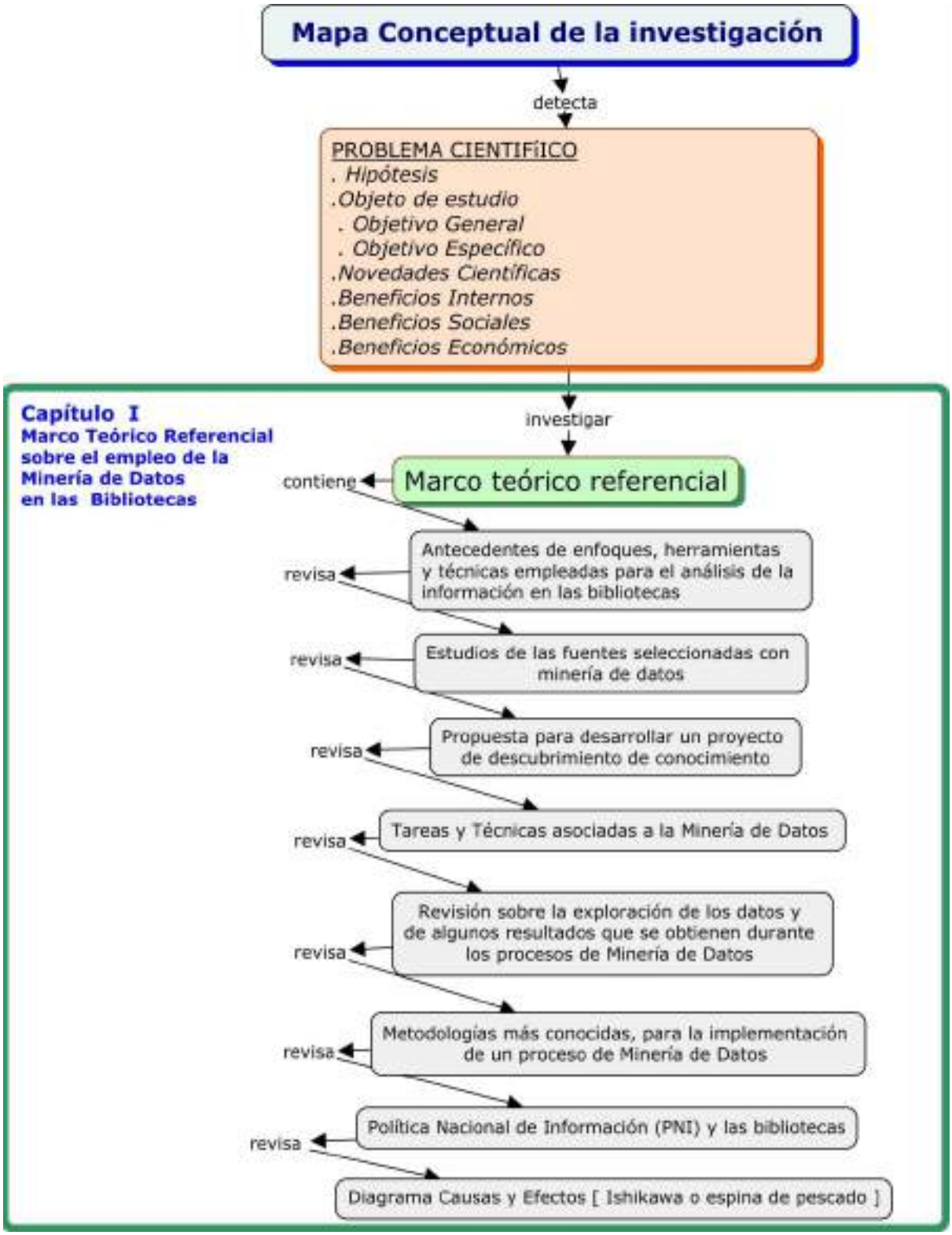


Figura 2. M**apa Conceptual de la investigación** para
el estudio de los Datos Bibliográficos con el empleo de la Minería de Datos
Fuente: elaboración propia.

Conclusiones parciales

El estudio del marco teórico referencial permitió llegar a las siguientes conclusiones:

1. Las técnicas de minería de datos constituyen técnicas de análisis muy fáciles de implementar (árboles de decisión, K vecinos más cercanos y las técnicas de visualización como los mapas auto-organizados de Kohonen y las tablas multidimensionales) permiten trabajar con muestras y ofrecen excelentes resultados.
2. Se reconoce la importancia y necesidad de determinar y gestionar el conocimiento necesario para contribuir al alineamiento estratégico por medio de las tareas de modelación *(*clasificación, predicción y segmentación) que permiten su rápida ejecución y comprensión.
3. La investigación mostró que el proyecto de descubrimiento de conocimiento de Cabena es el más adecuado y sencillo de implementar por constar con 5 procesos para la realización de la minería de datos, proporcionando una excelente calidad de la información en las bases de datos que se crean.
4. Se presenta como herramientas el software RapidMiner V4.6 que permite la ejecución de la minería de datos y el WinIDAMS13-SP software dedicado a la estadística, ambos presentan facilidades de ejecución y son softwares libres.

Capítulo II

Capítulo II

Desarrollo de una metodología para el estudio de los datos bibliográficos con el empleo de la minería de datos

Este capítulo presenta una metodología que ha sido desarrollada, para dar cumplimiento al objetivo general de esta investigación.

Como se definió el objeto de estudio para esta metodología es la información de las bases de datos bibliográficas. Este tipo de información está compuesta por los metadatos de cada documento que se registra en la biblioteca. Los datos en su gran mayoría es información textual y por esa razón solo se seleccionaron los campos que se consideraron como variables significativas para el estudio.

2. Requisitos que debe cumplir la información para entrar en el estudio.

Antes de comenzar la aplicación de la metodología se deben cumplir los siguientes requisitos:

- Seleccionar una base de datos bibliográfica;
- Esta base de datos debe estar creada por el propio sistema de gestión bibliotecario, porque los resultados del proceso de minería de datos se cruzarán contra el mismo sistema de gestión;
- La cantidad de registros de la base de datos debe ser grande, como para poder calcular muestras considerables;
- La información de los campos debe aportar patrones al ser tratada con la minería de datos;
- La información de cada campo debe aportar algún tipo de estadística, aunque la mayoría de los campos recoge información de tipo nominal y no numérica;

La preparación de la información debe, principalmente, estar basada en su estandarización y la preparación de la estructura de la base de datos, para poder aplicar la minería de datos.

Los registros de esta base de datos se dividen según la fórmula que se utiliza cuando se conoce el tamaño de la población (Pickers, 2015) y se forman diferentes tamaños de muestras. Estas muestras sirven para demostrar que los casos donde las computadoras no cuenten con la tecnología adecuada, para analizar una base de datos grande, pueden hacer estudios con muestras de la base de datos.

2.1 Metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencias y Técnicas

La metodología desarrollada para el estudio de datos bibliográficos con el empleo de la minería de datos se presenta en la figura 3.

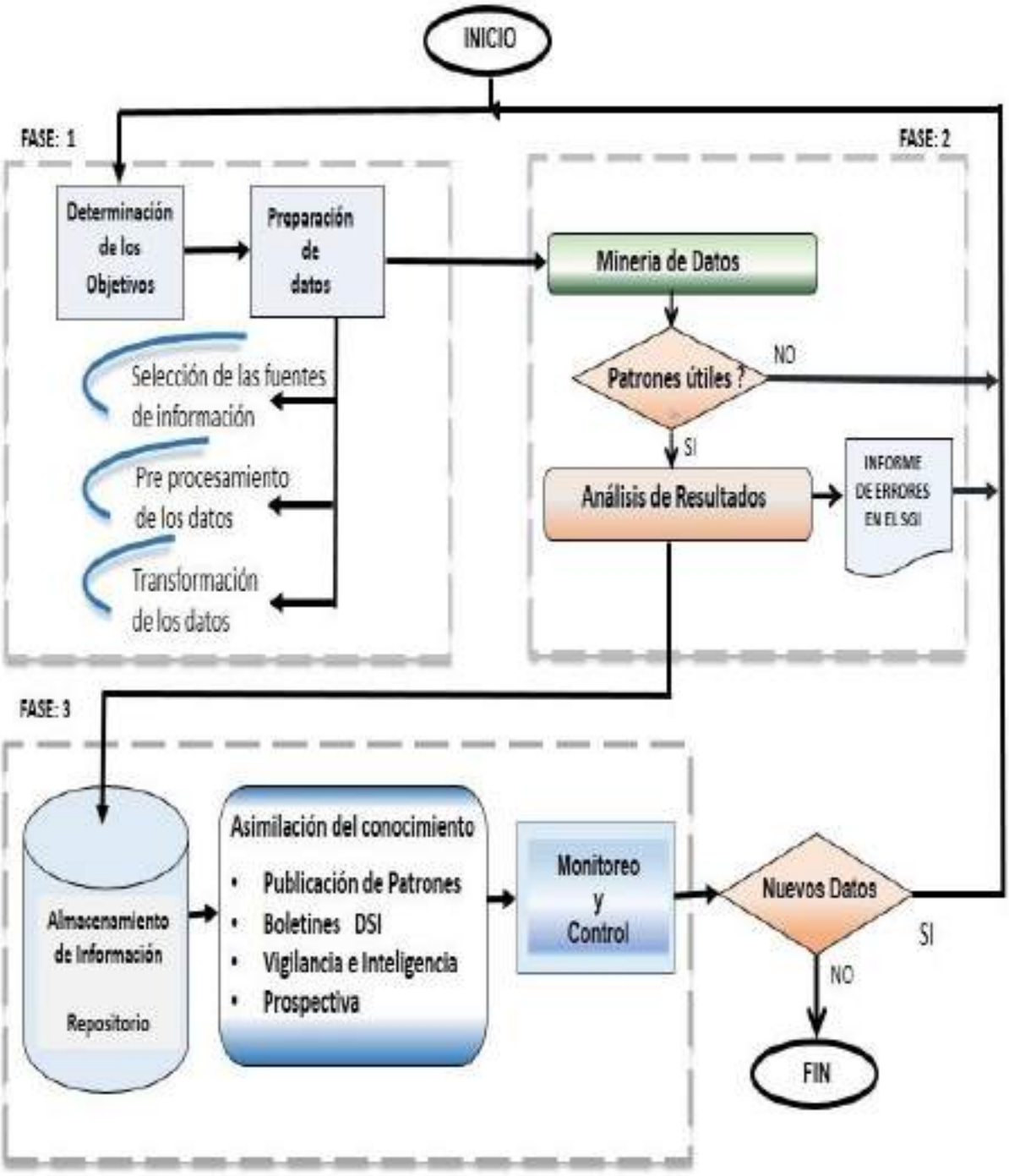


Figura 3.- Metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencias y Técnicas.
Fuente: Elaborado por la autora

2.2 Descripción de la Metodología presentada

La metodología desarrollada consta de 3 fases y cada una está integrada por procesos de la siguiente manera:

Fase 1

- Proceso de Determinación de los objetivos.
- Proceso de Preparación de datos.

Fase 2

- Proceso de Minería de datos.
- Proceso de Determinación de los datos útiles.
- Proceso de Análisis de los resultados.

Fase 3

- Proceso de almacenamiento de la información en el repositorio.
- Proceso de Asimilación del conocimiento.
- Proceso de Monitoreo y control.

2.2.1 Fase 1: Preparación de los datos

- *Determinación de los Objetivos del estudio*

Este proceso es el punto de partida para conocer las posibles amenazas y ventajas que pueden influir sobre el buen funcionamiento de la investigación, además se definen los objetivos que se quieren lograr al finalizar la misma. En este caso estos aspectos fueron presentados en la figura 1, donde se muestran cuatro grandes causas que son determinantes en la biblioteca, como son los datos que van a entrar al estudio, el hardware, los softwares que se van a utilizar en el estudio de los datos y las personas que van a hacer estos estudios, que deben poseer el conocimiento, para poder desarrollar todo el proceso hasta el final.

Como cuestión inicial es necesario definir los datos que se tienen para el estudio y también hay que definir las variables que se van a estudiar, porque los resultados de la investigación no serán los mismos si la información no es del tipo bibliográfica. Los objetivos dependen primeramente de la información a analizar, porque dependiendo del tipo de información, de la calidad, la cantidad y el formato de la base de datos que se tiene para hacer el estudio, se determina si existen condiciones para investigar y los objetivos que se pueden alcanzar.

Por otra parte, el hardware juega un papel importante debido a que permite asegurar el funcionamiento de los algoritmos que se utilizan y de igual forma las competencias de las personas permite desarrollar aspectos tales como:

Implementar, ejecutar, obtener los resultados así como, hacer una correcta interpretación de los mismos y finalmente presentar los resultados a los decisores, para que se cumplimenten los objetivos propuestos.

Por lo tanto, considerando el diagrama causa-efecto definido para cada investigación, se trazan los objetivos, se escogen posibles variables, se escogen los posibles algoritmos a ejecutar y se pasa a realizar el proceso de Preparación de datos.

- *La Preparación de los datos*.
Aquí se ejecutan todas las acciones de preparación de datos para con posterioridad comenzar con el *Pre procesamiento de los datos.* Es decir la preparación de los datos es considerado, como un proceso donde se invierte más del 50% del tiempo de trabajo, que se necesita para desarrollar toda la metodología. Esta preparación incluye el cambio de estructura y de formato de la base de datos, para llevarlo a otro software donde será pre procesado.

Para lograr el pre procesamiento de los datos también es necesario utilizar herramientas digitales, que ayuden a resolver los problemas que tiene la información, como son los campos vacíos, o campos con información sin valor para el estudio, registros dobles, falta de estandarización en la información, errores de digitalización, entre otros, (García et al., 2016).

Aquí se usaron 3 software para preparar la información, uno - EditPad V3.5.1 ES y el otro Procite V5.
El primero fue utilizado porque este editor de texto fue el que leyó la información original sin distorsionar los acentos y otros símbolos o caracteres especiales, además de hacer las modificaciones que se pidieron. El segundo software nos permitió preparar la estructura de los datos, porque convirtió la estructura original en forma de tuplas, es decir, se utilizó el software Procite V5 que convirtió cada registro en una fila y cada campo de los registros los separó con punto y coma. Toda esta nueva estructura de los datos se guarda en un fichero de texto (TXT), para que pueda ser realizada la importación al tercer software, Microsoft Excel.

Una vez importada la información a Microsoft Excel, se pueden modificar grandes volúmenes de información de forma rápida y fácil, permite buscar errores y

rápidamente sustituirlo por un nuevo valor que se le ordene, permite eliminar y agregar columnas y filas, además que se puede guardar la información en diferentes formatos, en especial en formato CVS o XLS, que son formatos que admiten el software de minería de datos seleccionado para extraer los patrones. Microsoft Excel trabaja la información en forma de tabla, donde cada columna representa una variable y una fila representa un registro de la base de datos original.

Una vez que la información se encuentra en Microsoft Excel, correspondiendo columnas con variables que se van a estudiar y filas con número de registros que va a tener el fichero, ya se encuentran los datos listos para empezar, con el proceso de estandarización y revisión de la información de cada uno de los campos.

El proceso de estandarización consiste en adoptar una forma única de escribir un término que se repite más de una vez, por ejemplo el nombre de un autor, el nombre de una revista, los nombres compuestos, que en ocasiones son escritos de forma completa o de forma abreviada con las iniciales, consideradas estas dos formas de escritura como dos personas diferentes para el caso de los buscadores, sin embargo ambas escrituras corresponden a la misma persona y debe estandarizarse.

Terminada la estandarización de la información, se comienza con la eliminación de los registros repetidos y después se decide sobre los campos que están vacíos, es decir, se sustituye por algún carácter especial que los identifique al final del proceso de minería de datos o se decide si son eliminados estos registros definitivamente.

Finalizado el pre procesamiento de la información, se reporta al departamento o al personal encargado del mantenimiento de esta fuente de información. Terminada estas correcciones, directamente se ha logrado mejorar la recuperación de la información, es decir, la reparación de los errores en la base de datos fuente, permite mayor eficiencia en la recuperación del sistema de gestión bibliotecario (SGB) (Nicholson, 2003 b).

- *La Transformación de los datos*

Aquí se analiza la cantidad total de registros y se conforman los distintos ficheros de pruebas (muestras) que entrarán al estudio, además que estarán conformados solo

con las variables más influyentes, en relación al algoritmo de minería de datos que se quiere ejecutar en cada prueba, estos son:
Árboles de decisión (clasificación y regresión), los k vecinos más cercanos con las técnicas de visualización (mapas auto-organizados de Kohonen) y las tablas multidimensionales.

La decisión de particionar la cantidad de registros total y crear ficheros de muestra, son la solución para investigaciones como esta, donde se cuenta con computadoras de muy bajas prestaciones tecnológicas.

En esta investigación se crearon los distintos ficheros de prueba y para esto se tuvieron en cuenta dos aspectos:

1) la selección de las variables que tienen valor para el estudio que se va a realizar y se desechan aquellas que no tienen valor;
2) se escogieron diferentes cantidades de registros para conformar los ficheros que serán estudiados.

Por último, se definen los campos que son de texto y los campos numéricos, cual campo puede ser una variable dependiente y cual no, dependiendo del algoritmo según el caso que se va a ejecutar y por último se decide la cantidad de registros que van a entrar en cada estudio.

El cálculo del tamaño de la muestra (Pickers, 2015), se realiza a través de la fórmula (1) que plantea, que conociendo el tamaño de la población se puede calcular el tamaño de la muestra.

$$n = \frac{N \times Z_a^{\,2} \times p \times q}{d^2 \times (N-1) + Z_a^{\,2} \times p \times q}$$

(1)

Donde:
N = tamaño de la población,
Z = nivel de confianza, (Un intervalo de confianza de 95% significa que los resultados de una acción probablemente cubrirán las expectativas el 95% de las veces)

P = probabilidad de éxito, o proporción esperada,
Q = probabilidad de fracaso,
D = precisión (error máximo admisible en términos de proporción).

Por último, toda esta transformación de los datos, que respetó el tipo, cantidad, calidad y formato de los datos fue guardada, para hacer la importación al software seleccionado y ejecutar los algoritmos de minería de datos.

En resumen, en la Fase1 se determinan los objetivos de la investigación, se debe considerar previamente la calidad, cantidad y formato de la información que se tiene para hacer el estudio, como ya se explicó y de igual manera, dependiendo de los objetivos que se quiere lograr y de la máquina que se tiene para hacer el estudio, se escogen las herramientas y los algoritmos que van a ser utilizados en la Fase 2.

La forma de lograr el éxito en la determinación de los objetivos y la preparación de datos difiere totalmente entre diferentes tipos de datos y diferentes tipos de investigaciones, y estas diferencias son las que hacen los diferentes aportes en cuanto a:

- la forma de extracción de los datos que se quieren estudiar;
- la forma de preparación de los datos en relación a estandarización, campos vacíos; tipo de información (numérica o texto), tamaño de la muestra;
- la separación de la información sensible, de la información no sensible, para poder lograr los objetivos específicos y generales definidos en la investigación;
- el nuevo formato de las muestras para su importación a las herramientas de minería;
- el informe de los errores detectados para arreglar la base de datos;
- el mejoramiento en la recuperación de la información en el SGB;

Todo esto demuestra que la Fase 1 es determinante para el buen desarrollo de las dos fases restantes, cada una de estos procedimientos generales van a procedimientos mucho más específicos para poder lograr los objetivos. Estos procesos se consideran específicos, porque varían en cada investigación en dependencia de los objetivos propuestos, del conjunto de datos seleccionado para el estudio y del modelo analítico o algoritmo seleccionado, para desarrollar en cada investigación.

2.2.2 Fase 2: Aplicación de la minería de datos

- *Proceso de Minería de datos*

Aquí se importa la información que se ha preparado para los dos software seleccionados, el RapidMiner V4.6 y el software WinIDAMS13-SP. Con el RapidMiner V4.6, se procede a la ejecución de los algoritmos de minería de datos previamente escogidos, el primer algoritmo que se ejecuta es un árbol de decisión y el otro algoritmo es el SOM. Cuando se extraen todos los resultados propuesto con RapidMiner V4.6, se pasa a trabajar con el software estadístico WinIDAMS13-SP y se obtienen las tablas multidimensionales, barras 3D y gráficos de correlación. Todos los resultados obtenidos, se guardaron e identificaron para luego entrar en el proceso de Análisis de resultados.

Se revisó la calidad de los resultados basado en la literatura, que plantea que dependiendo de las características de un modelo de minería de datos, existen diferentes criterios que se deben cumplir (Microsoft, 2014), como son:

1. La separación de los datos en conjuntos de entrenamiento y de prueba con el fin de probar la precisión de las predicciones.
2. El uso de varias medidas de validez estadística para determinar si existen problemas en los datos o en el algoritmo seleccionado.
3. La revisión de los resultados del modelo de minería de datos, para determinar si los patrones detectados cumplen con los objetivos planteados.

Para comprobar esto se procede al Análisis de resultados, donde se revisan las métricas de la minería de datos que son: la precisión, la confiabilidad y la utilidad (Microsoft, 2014). Algunos softwares de minería de datos calculan por sí mismo estas métricas, pero en otros casos es necesario aplicar fórmulas, para calcular la precisión, se emplea la fórmula (2).

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad \textbf{(2)}$$

Donde:

Relevant documents = documentos relevantes
Retrieved documents = documentos recuperados

Estas métricas proporcionan medidas objetivas durante el análisis de los resultados, y por ejemplo se utilizan para evaluar la confiabilidad de los datos en los análisis predictivos (Microsoft, 2014), es decir la confiabilidad evalúa la manera en la que se comporta un proceso de minería de datos en conjuntos de datos diferentes.

La precisión es otra medida que indica hasta qué punto existe una correlación entre un resultado, con los atributos de los datos que se están estudiando, mientras que la utilidad incluye diferentes métricas que indican si los patrones logrados son información útil (Microsoft, 2014), porque de no ser útiles estos patrones, se debe empezar desde el inicio a buscar que debe modificarse y comenzar un nuevo proceso.

Otro de los procedimientos que se realizan durante el análisis de resultados consiste en hacer una validación cruzada entre los patrones de la tabla *Text View* creada por RapidMiner V 4.6 durante la ejecución de un árbol de decisión contra la recuperación de esos mismos patrones en el SGB, mientras que para el caso del SOM, el propio software crea un mapa que permite conocer los valores de las células que están fuera de su zona de clusterización, para luego buscarlos en el SGB y conocer que provocó este comportamiento.

Este tipo de validación cruzada entre patrones de la minería de datos y recuperación del SGB, permitió rápidamente demostrar la utilidad de los patrones, porque de ser correcta la recuperación de la información por el SGB, queda demostrado la utilidad del patrón, además del buen funcionamiento del SGB, de lo contrario si la recuperación de la información es incorrecta, entonces con estos errores se crea un “Informe de Errores” y se envía a los responsables del mantenimiento de la base de datos, para reparar los registros con mala recuperación en el SGB y después de la reparación nuevamente se ejecuta el algoritmo, para conocer si existe cambio en el comportamiento de ese patrón y en el SGB.

2.2.3 Fase 3: Proceso de almacenamiento

- Proceso de Almacenamiento de la información en el repositorio

Este proceso se encarga de guardar en el repositorio creado con la herramienta Eprint, los registros que están relacionados con los patrones de la minería de datos y que fueron validados con el SGB, es decir, los trabajos recuperados por el SGB que no tienen errores, son validados con base de datos de patentes y otras bases de datos internacionales, ejemplo la web de Word Intellectual Property Organization (WIPO

Patenscopea). EL objetivo de esto es identificar trabajos afines con calidad y debidamente registrados con los patrones de la minería de datos y toda esa información recuperada pasarla a formar parte de la información del repositorio, para que sirva en estudios posteriores.

Este tipo de búsqueda en otras bases de datos que no pertenecen a la biblioteca de Ciencia y Técnica, estuvo incentivada porque los patrones de clusterización que aportó la minería de datos infieren que patrones con mayor clúster, es porque también existe un mayor número de metadatos que forman parte de ese clúster y a su vez un mayor número de trabajos que los investigadores nacionales están desarrollando en esa temática donde pertenece el metadato, por lo tanto se debe conocer el trabajo de los investigadores extranjeros, que también investigan sobre esos mismos campos de investigación, razón que justifica la búsqueda en sitios de patentes y bases de datos internacionales, para poder comparar estudios similares entre investigadores nacionales y extranjeros.

Este paso es la segunda forma de añadir valor al repositorio. Este repositorio debe guardar la información resultante de la minería de datos, además de toda la información recuperada de estas validaciones cruzadas, primero con el SGB y luego con la búsqueda en sitios de patentes y bases de datos internacionales, ya que esta información será utilizada en el proceso siguiente.

- *Proceso de Asimilación del conocimiento*

El repositorio creado es una de las novedades científicas de este trabajo por su contenido y entre sus objetivos está el crear productos y servicios que faciliten la asimilación del conocimiento.

Entre los servicios a ofertar con el repositorio está:

- Publicación de Patrones.
- Boletines de Difusión Selectiva de Información (DSI)
- Vigilancia e Inteligencia
- Prospectiva

Una vez guardada toda la información correctamente y de que este repositorio esté en línea, ya puede ser consultado por todos los interesados, este paso ahorra tiempo

tanto al cliente como a la biblioteca, porque no va a necesitar un especialista a tiempo completo en estas funciones de búsquedas, a no ser que el cliente lo solicite por determinadas razones.

A través de las consultas por parte de los clientes al repositorio puede surgir la difusión de información por demanda (Castillo, 2013), que consiste en la iniciativa por parte del usuario, de solicitar una información concreta como ha venido ocurriendo. Por lo general cuando un usuario ha pedido alguna información al centro, lo que ocurre es que un especialista hace las búsquedas en las bases de datos que ha creado el propio centro y también sobre internet y sobre bases externas compradas y con todos estos resultados se confeccionan los Boletines de DSI, que en ocasiones son enviados en forma digital por email y en otros casos se envían en formato de papel a un correo postal.

Sin embargo, la forma propuesta, es que el mismo cliente consulte el repositorio que está en línea y que decida si le interesa esa información o necesita algo diferente. Esta última propuesta es conocida como difusión activa (Castillo, 2013), porque el centro crea y brinda productos documentales que considera útiles, como es el caso de los patrones de la minería de datos, además de toda la información recuperada del SGB y de bases de datos internacionales, por ser información que está asociada al comportamiento de los patrones encontrados. Esto es considerado como un mejor uso de las tecnologías de la información y las comunicaciones, para lograr una mayor eficiencia del trabajo de la biblioteca.

De igual forma el repositorio está preparado para formar parte de las fuentes que pueden ser usadas en un proceso de vigilancia, ya que ha sido creado con la información proveniente del proceso de minería de datos, es una herramienta con un alto valor agregado. Este repositorio también puede ser utilizado en estudios prospectivos, que se define como: las *"tentativas sistemáticas para observar a largo plazo el futuro de la ciencia, la tecnología, la economía y la sociedad, con el propósito de identificar las tecnologías emergentes que probablemente produzcan los mayores beneficios económicos y sociales"* (Medina, 2015).

La prospectiva permite identificar grandes tendencias de evolución, que de acuerdo a la información bibliográfica que se estudió con la minería de datos, pueden ser sobre cantidad de investigaciones por metadatos, cantidad de investigaciones por autores, tipo de tecnología utilizada en determinada investigación, autores extranjeros o países involucrados en determinado tipo de investigación, entre otros, lo que permite proponer recomendaciones, instrumentar sistemas de seguimiento, y obliga a crear un sistema de boletines periódicos para informar a los interesados.

- Monitoreo y control

El monitoreo y control persigue tres objetivos fundamentales:

- Decidir si se puede comenzar un nuevo estudio basado en la cantidad de nuevos registros incorporados en la base de datos. Esta decisión se toma en base a la cantidad menor de una muestra, que quedó definida a través de la fórmula de Pickers (fórmula 1). Una cantidad menor de este cálculo no justifica comenzar un nuevo estudio.

- Monitorear el nivel de satisfacción del personal bibliotecario sobre la metodología desarrollada y aplicada. Se recogió en Google a través de un formulario de 5 preguntas que contempla los aspectos fundamentales del proceso de aplicación de la Metodología creada. El análisis de estos resultados se hizo dirigida a la frecuencia de las variables y su porcentaje, utilizando el software estadístico IBM SPSS Statistics v.21 (anexo 30).

- Monitorear el nivel de satisfacción del cliente sobre los servicios bibliotecarios se recogió a través de una encuesta de 8 preguntas, de las cuales 2 preguntas son directamente sobre la metodología creada y 1 pregunta sobre las necesidades del cliente que la biblioteca no tenga contemplada. Para este análisis se utilizó la Técnica de Iadov (anexo 31).

Conclusiones parciales

1. La metodología propone tres fases que determinan los objetivos de la minería, el tratamiento de la información y por último el inventario de conocimiento guardado en el repositorio institucional.

2. Los algoritmos utilizados en el desarrollo de la metodología permiten lograr los diferentes resultados y la métrica de confiabilidad, que confirman los resultados logrados.

3. Se comprobó el nivel de satisfacción de los servicios bibliotecarios por medio del software estadístico IBM SPSS Statistics v.21 y del cliente externo por medio de la técnica de Iadov (índice de satisfacción grupal) sobre la metodología empleada y ambos resultados se considera máxima satisfacción y por tanto, se interpretan como una valoración positiva.

Capítulo III

Capítulo III

Aplicación de la metodología para el estudio de los datos bibliográficos en la Biblioteca de Ciencias y Técnicas

3. Resultados y análisis de la aplicación de la metodología

El capítulo III presenta los resultados obtenidos, y su análisis durante la aplicación de la metodología, que fue ejecutada paso a paso, con la información bibliográfica que se conserva en la Biblioteca de Ciencias y Técnicas, ubicada en la entidad.

3.1 Fase 1. Resultados obtenidos en la Preparación de los datos

El proceso de Preparación de los datos se logró con el desarrollo de los tres procedimientos fundamentales declarados en la metodología:

a) Selección de la fuente de información

La selección de esta fuente de información se determinó por ser la información bibliográfica que conforma el catálogo que utiliza el Sistema Gestor Bibliotecario (SGB). En esta base de datos están recogidas las publicaciones de cuatro segmentos de la ciencia: Ciencias Agropecuarias, Ciencias Técnicas, Ciencias Sociales y Ciencias Biomédicas y el tipo de información que tiene cada segmento guardado en la base de datos son los metadatos de: tesis (T), premios académicos (PAC), publicaciones seriadas (S) y los manuscritos depositados (MD).

b) *El Pre procesamiento de los datos*, consistió en la limpieza y reestructuración de la información. Este primer proceso de tratamiento de la información se realizó utilizando el editor de texto EditPad V3.5.1 ES. Con este editor se eliminaron campos que guardaba información que se considera ruido para el estudio. Ejemplos de este tipo de información son: el campo con etiqueta <935> que recogía el nombre de la operadora que había introducido el registro o el campo <5>, que recogían la fecha y hora sobre la captación y modificación de cada registro, entre otros. Toda esta información fue retirada, por no aportar valor a la investigación.

c) *La Transformación de los datos,* proceso obligatorio de cambio de estructura porque la información se encuentra en forma columnar y hay que llevarla a tuplas. Para lograr esto se utilizó el software Procite V5. Este software permite la importación de la información a través de una plantilla, donde hay que relacionar las

etiquetas que tiene la información que se está estudiando, con las etiquetas que utiliza el software en su plantilla de importación. Después de importada la información a Procite V5, se continuó con la limpieza de la información, aún quedaba información que creaba ruido, y se procedió a crear la nueva estructura de los datos (tabla 3).

Campo	Descripción del campo
Autor:	Nombres y Apellidos del autor/es de la publicación.
Autor/es secundario/s:	Nombres y Apellidos de existir coautor/es.
Filiación:	Lugar donde pertenece el autor de la publicación.
Titulo:	Título de la publicación.
Titulo Traducido:	Titulo traducido.
Edición: y fecha:	Lugar donde se editó la publicación y fecha de edición.
ISSN:	No. de registro que identifica la publicación.
BNCT:	No. de localización en de los estantes de la biblioteca.
Revista:	Nombre de la Revista que contiene la publicación.
TD:	Tipo de documento.
Year:	Año de publicación.
Segmento:	Segmento de la ciencia al cual pertenece la publicación.
Idioma:	Idioma/s en que han sido traducida la publicación,
Resumen:	Abstracto de la publicación sobre su contenido y objetivos.
Keyword1, Keyword2, Keyword3, Keyword4, Keyword5, Keyword6, Keyword7, Keyword8:	Contienen los metadatos, por los cuales se puede recuperar las publicaciones en el sistema de gestión.

Tabla 3. Campos de la base de datos que se está estudiando.
Fuente: elaboración propia.

Concluido el trabajo con el software Procite V5, se exportó la información a formato .TXT, con el objetivo de completar la transformación de la información columnar a tuplas y con registros muchos más limpios y ya listos para importarlos en el software de Microsoft Excel.

Una vez que cada registro se convirtió en una fila y cada campo se convirtió en una columna de la tabla de Microsoft Excel, empezó otro tipo de limpieza dirigida a:

- Los campos *Keyword7* y *Keyword8* fueron retirados de la base de datos por tener más del 70% de los registros vacíos, es decir sin información y por lo tanto se consideraron como información no útil para esta investigación.
- Los campos considerados como información útil para la investigación, pero que presentaban muchos registros vacíos o sin información se decidió rellenarlos con un carácter que permitió identificarlos dentro de los resultados, en este caso se utilizó el signo de interrogación (?).

- Se eliminaron los registros que tenían vacíos los campos: *Segmento* y *Autor*, porque son variables útiles, pero sin información no son útiles.

- Separación de la información relevante, de la información no relevante según el algoritmo de minería de datos a ejecutar.
  En este caso los Autor/es secundario/s, *Filiación, Titulo Traducido, Edición: y fecha, ISSN, BNCT, Resumen* fueron retirados por no ser considerados información relevante en este momento para la investigación.

Todos estos pasos se repitieron para los cuatro segmentos de la Ciencia (Ciencias Agropecuarias, Ciencias Técnicas, Ciencias Sociales y Ciencias Biomédicas) y cuando las cuatro tablas en Microsoft Excel quedaron preparadas, se procedió a rectificar los errores ortográficos, la estandarización de la información, la sustitución de nombres extremadamente largos por siglas o abreviaturas, como en el campo *Revista*. Vale aclarar que todos estos pasos, se logran trabajando directamente con las herramientas EditPad V3.5.1 ES, WordPad 2009, Procite V5 y Microsoft Excel, herramientas utilizadas para reparar la información. En dependencia de los tipos de errores y la estructura de cada información a preparar para los procesos de minería de datos, se seleccionan las herramientas a utilizar. Ejemplo: si tiene la información en tuplas, entonces no es necesario usar el Procite V5.

Esto implica, que la cantidad de herramientas a utilizar en la Fase 1, depende de las características que tenga la información. Esta etapa de Preparación de los datos, es costosa según algunos autores, porque también se necesita de un personal preparado que sepa determinar los objetivos y a partir de esto, preparar la

información, seleccionar los algoritmos que se van a ejecutar y los campos que pueden aportar patrones.

También se debe contar con una computadora que disponga de suficiente espacio en el disco duro, para ir salvando los resultados de cada limpieza y transformación, de una memoria RAM grande y una suficiente velocidad en el micro procesador, para manejar grandes volúmenes de información, porque todos estos pasos intermedios generan nuevas tablas que es necesario guardarlos, por si se necesita ir atrás y reconfigurar los datos nuevamente. Esto es muy frecuente que suceda cuando se están haciendo pruebas y se decide reevaluar la información en relevante o no relevante, o en otro caso que se detecten nuevos errores.

Concluida la limpieza, selección de las variables y unificación de los cuatro segmentos de la Ciencia en una misma tabla de Microsoft Excel, se guarda la información en formato .TXT, en formato .CVS y en formato .XLS.
Por tanto, la primera estructura creada para entrar al proceso de minería de datos se encuentra distribuida por columnas, con los siguientes campos y en el siguiente orden:

Segmento, Year, Revista, TD, Idioma, Keyword1, Keyword2, Keyword3, Keyword4, Keyword5, Keyword6 y Autor

Tabla 3A. Campos seleccionados de la base de datos en estudio.

Estos tres nuevos formatos (.TXT, .CVS y .XLS) se prepararon de acuerdo a los requerimientos que exigen las distintas herramientas de minería de datos para hacer la importación y comenzar con el desarrollo de la Fase 2.

Esta fase 1 permitió detectar y corregir los errores de la información, que fueron introducidos durante el periodo de captación, elemento que ofreció dos ventajas: una mejora considerable en el proceso de recuperación del SGB al trabajar con una base de datos a la cual se le habían reparado los errores detectados y otra contar con una información preparada para importar a las distintas herramientas de minería de datos.

Como se explicó, es necesario para computadoras con bajas prestaciones tecnológicas, hacer el cálculo de las diferentes muestras a procesar, así, con el conocimiento del tamaño de la población total, se hizo el cálculo, con el empleo de la fórmula 1 (capítulo II), esta fórmula también se puede calcular desde un sitio web (SurveyMonkey, 2018), donde se obtiene el resultado, después de entregarle los datos siguientes:

Tamaño de la población (10492)

Nivel de confianza (95%)

Margen de error (5%)

Como resultado del cálculo se obtuvo que las muestras deben tener 371 registros aproximadamente, por tanto, se creó un fichero con 400 registros para redondear el resultado de la fórmula y se crearon otros dos ficheros, uno muy por debajo del cálculo, con 100 registros y otro con los 10492 registros total y se le aplicaron todas las técnicas de minería de datos previamente seleccionadas. Los resultados del estudio de estos 3 ficheros fueron similares, por lo tanto las computadoras de bajas prestaciones no afectarán los resultados de los estudios si utilizan los ficheros de muestra.

3.2 Fase 2. Resultados obtenidos en la Aplicación de la minería de datos

En esta fase se agrupan dos procesos, uno, la *minería de datos* y dos, el *análisis de resultados*, este último proceso comprende la validación cruzada. En esta fase 2 estuvieron involucrados cuatro softwares como fueron: Microsoft Excel (ficheros de datos para exportar), Rapid Miner V 4.6 (software para minería de datos), WinIdams V.13 SP (software de minería y estadística para análisis de los datos) y Altair (SGB), (software gestor bibliotecario utilizado para la validación cruzada de los resultados de la minería).

La ejecución de la fase 2 comenzó con la importación de los datos que se encontraban en Excel con formato .CSV a Rapid Miner V 4.6 y después a WinIDAMS13-SP.

Con Rapid Miner V 4.6, se explican los casos estudiados con las diferentes muestras preparadas para encontrar patrones y continúa con WinIDAMS13-SP, que son los casos estudiados con las diferentes muestras preparadas para WinIdams V.13 SP, también para encontrar patrones y para demostrar estadísticamente la fortaleza de las variables estudiadas.

RAPID MINER V 4.6

Rapid Miner V 4.6 es una herramienta que necesita que los datos le sean importados con determinadas condiciones, estas condiciones fueron cumplidas con la creación de los ficheros en Microsoft Excel con formato .CSV.

Todas las ejecuciones que se realizaron en Rapid Miner V 4.6, fueron para obtener los árboles de decisión, para esto se personalizó la herramienta seleccionando un operador, que importa la información desde una hoja de cálculo de Microsoft Excel, **Anexo 1** (panel de la izquierda) y un segundo operador que es el algoritmo encargado del análisis de la información y de crear el árbol en estos casos.

En el primer operador se personalizaron los parámetros file excel, label-column y id-column de la siguiente manera: file excel es el parámetro donde se declara la dirección donde se encuentra la hoja de cálculo con la información en la computadora y los otros dos parámetros (variables), son las posiciones de las columnas en la tabla de Microsoft Excel que tienen la información que se quiere analizar, es decir estas dos columnas o campos son las variables que determinan los nodos y las hojas del árbol que se va a obtener.

En este primer caso se tomó por label-column = 7, que representa el campo keyword2 de la tabla de Microsoft Excel y el id-column = 2, que representa el campo Year (Año) de la tabla Excel, que pertenece a un fichero de muestra con 400 registros (anexo 1).

Después de la importación de los datos, el software creó automáticamente una tabla sensible nombrada como ExampleSet y con la cual se puede interactuar. Esta tabla nombrada Meta Data View, presentó los campos por filas y brindó una estadística y análisis preliminar de los datos por columnas, información muy útil. Estos resultados estadísticos se refieren al número de ocurrencias de la información dentro de cada campo (clasificación) y a partir de estos resultados, ya se puede aplicar cualquier algoritmo y gráfica de los que brinda Rapid Miner V 4.6 a los datos (anexos 2 y 3).

Estos resultados que brinda la herramienta de forma automática, permite revisar toda la información que ha procesado el algoritmo y lo muestra organizado en forma de campos clusterizados, además se encuentra entre paréntesis al lado de cada campo

el número de ocurrencias o frecuencias total de cada clúster. Si aún se quiere detallar más la información, sobre la clusterización de algún dato en específico, bastará poner sobre la información que se desee el puntero del mouse y una ventana emergente ampliará los detalles de esa información, porque son tablas sensibles creadas para resaltar los detalles en cada uno de los campos de la tabla, (anexo 2 y 3).

Cuando la información que muestran estas ventanas emergentes excede al tamaño de la pantalla, como es el caso del anexo 3, que se pierde la visibilidad completa de los datos, entonces se pueden obtener los datos con el Text View, (anexo 8). La ventaja de las ventanas emergentes del Meta Data View consiste en que se puede detectar cualquier clasificación que se desee, ejemplo en los autores, (anexo 2 y 3).

Cáceres Lóriga; Fidel Manuel,
Fajér Ávila; Victor,
Pérez Ramírez; Miguel,
Zayas Molina, Roberto,

Todos con 3 publicaciones cada uno, según esta muestra de datos de 400 registros, de la cual procesó 399.

Esta herramienta sirvió para procesar información tanto del tipo textual como numérica, y la estadística que ofrece, está basada en la clusterización del conjunto de datos que se está procesando. En la parte superior de esta ventana de resultados, (anexo 2), se puede acceder a esta información a través de un sistema de tres pestañas nombradas: la *Meta Data View*, la *Data View* y el *Plot View*.

La pestaña *Meta Data View* describió el tipo de información que se encuentra en cada columna, se muestra la información de cada campo y se agrega la información estadística de cada atributo. El orden de esta tabla se recoge en columnas nombradas: Type (Tipo), Name (Nombre), ValueType (Valor Tipo), Statistics (Estadística), Range (Rango) y Unknown (Desconocido), pudiendo agregársele más columnas como: Index (Índice), Construction (Construcción), Block Type (Tipo de Bloque) y Sum (Suma). Además, se resalta en la parte superior de la ventana una línea informativa con la cantidad de registros procesados y sus características. En este primer ejemplo, la información procesada fue de 399 registros, la cantidad de atributos especiales (2) y la cantidad de atributos regulares (10) (anexo 2).

Esta tabla interactiva también permite ventanas emergentes con información sobre la clusterización de cada uno de estos campos. Esta clusterización, antes de ejecutar cualquier otro algoritmo, prioriza la descripción del conjunto de datos multidimensional complejo, para ayudar a la interpretación de los datos, (anexo 2 y 3).

En la pestaña *Data View*, se muestran los datos con el mismo ordenamiento por columnas que la información trae de Microsoft Excel, además da la posibilidad de una ventana desplegable llamada *View Filter*, con la cual se obtuvieron otros valores de los atributos y etiquetas.

La información inicial de Rapid Miner V 4.6, que se muestra en el anexo 2, permitió crear informes de cada uno de los campos para futuros análisis, ejemplos: cantidad de publicaciones por Autor en la base de datos, cantidad de metadatos iguales dentro de la Keyword1, entre otros.

La pestaña *Plot View* es la tercera de estas ventanas y es por donde se accede a todos los algoritmos de visualización que ofrece Rapid Miner V 4.6, anexo 4.

Dentro de las ventajas de esta herramienta está la posibilidad de desarrollar todas las técnicas que recoge la *Tabla 1.Tareas de Modelación y Técnicas de Minería de Datos*, presentada en el *capítulo I*, y con ella obtener una visualización gráfica, que facilita la interpretación de los resultados.

Por lo tanto, después de estos primeros resultados que brinda el software en forma automática, la primera técnica consistió en visualizar el algoritmo encargado de obtener diferentes tipos de árboles como son el Árbol de Decisión, (anexos 5, 6, 7) y del árbol Chaid, anexo 9.

Con el algoritmo de Árboles de Decisión se hicieron varias pruebas y se obtuvieron diferentes gráficos, cada uno de ellos aportó diferentes patrones de comportamiento que se encontraban ocultos en los datos y que no podían obtenerse con una revisión manual, ni tampoco a través del gestor de información de la biblioteca. Los diferentes tipos de árboles se lograron, tomando diferentes juegos de variables o campos, que fueron escogidas para cada caso en particular y con diferentes tamaños de muestras, (anexos 5, 6 y 7).

Con las tareas de Segmentación y Agrupación se aplicaron técnicas de redes neuronales y visualizaciones, (anexo 4). Esta *técnica de minería de datos*, facilitó el análisis a través del algoritmo de la "distancia euclidiana", aquellos casos que se reagruparon por similaridad y que es imposible recuperar con un gestor de información.

Con la construcción de los árboles de decisión, se pudieron obtener dos tipos de información, una información gráfica sobre los árboles, (anexos 5, 6 y 7) y otra información textual o Text View, (anexo 8), que se tomó como ejemplo y que muestra las reglas de agrupamiento utilizadas en la construcción de ese árbol. Ambos resultados ilustran las diferentes distribuciones y asociaciones de los datos y permitieron estudiar donde se encuentra el mayor número de artículos dentro del conglomerado de datos, de acuerdo a la variable seleccionada (Ruiz et al., 2018).

En la Tarea de Predicción con la técnica del "k vecinos más cercanos" se lograron gráficos, donde se pueden apreciar los grupos que se forman producto de las reglas de clasificación por vecindad, mostrando el conjunto de los k prototipos más cercanos del patrón escogido a clasificar por series, todo este cálculo lo hace el software y en este caso las variables seleccionadas fueron Segmento y Autor, (anexo 4).

El Self Organizing Maps (SOM), gráfico basado en el campo Segmento (Ciencias Agropecuarias, Ciencias Biomédicas, Ciencias Sociales y Ciencias Técnica), detectó Autores, que se encontraban fuera de los clústeres formados por el criterio de los K vecinos más cercanos y como lo muestra en el margen superior de la gráfica. Este tipo de información permitió conocer que algo estaba situando a determinados Autores fuera de su Segmento, por tanto, también se analizaron a través del SGB, apoyado en el criterio de validación cruzada, como medida de comprobación, (anexos 15 y 16).

En el anexo 12 se muestra una matriz de correlación y la variable estudiada fue el campo Segmento. En la esquina superior izquierda, se observa una leyenda que identifica por colores los Segmentos, ejemplo el color azul como CAgrop, el verde a Biomed, el amarillo CTécn y el rojo a Csoc**.** Este análisis multivariables, muestra la densidad de cada una de estas variables. Por ejemplo, en este caso de la variable

Keyword6, convence, de que no es necesario el estudio específico de esta variable, aunque esto ya se había notado durante el proceso de limpieza de los datos, cuando se detectó la gran cantidad de campos vacíos en esta variable.

Todos estos gráficos se lograron con los tres tamaños de muestras, (100, 400 y por último los 10 492 registros total de la base de datos), y la salida gráfica de esta última muestra se ajustó para una visualización de 1000 registros solamente, porque en el caso de los 10 492 registros, es una información muy grande, para la computadora donde se ejecutó este estudio.

A los tres ficheros de muestras (100, 400 y 10492 registros) se le aplicaron los mismos algoritmos y los patrones encontrados fueron similares, además de que a cada fichero se le aplicaron 25 veces los algoritmos, porque ejecutando el mismo algoritmo, con el mismo fichero varias veces o solo haciendo pequeñas correcciones a la información, puede dar resultados muy diferentes en cada ejecución, ya que el software aprende sobre las relaciones que existen dentro de estos datos y como consecuencia de esto, se seleccionaron los resultados más relevantes para el estudio.

Otro resultado se presenta en el anexo 13, donde se muestra la creación de un árbol de decisión, tomando por id-column y label-column, las variables Segmento y Tipo de Documento respectivamente. Este árbol tiene una organización para los 4 segmentos, donde Csoci, CAgrop y Biomed están conformadas en su gran mayoría por publicaciones 'Seriadas' según se puede apreciar y también se comprobó en la propia base de datos. En el caso del segmento Biomed, se comprueba en la base de datos, que junto a las publicaciones Seriadas existen Premios Académicos y Tesis, en una cantidad significativa, por eso se justifica la coloración diferente en la hoja.

Para el caso del Segmento de CTéc el algoritmo hace un desglose por Revistas, para destacar cuales Revistas tienen publicaciones Seriadas y otras que tienen el campo de Revistas vacío y que han sido identificados con el signo de ‘?’, porque son Manuscritos Depositados.

Esta prueba fue realizada con una muestra de 400 registros de los 10492 y permite conocer cuatro resultados, (anexo 13)**.**

- Los Manuscritos Depositados (MD), tienen el campo Revista vacío y es correcta esta respuesta.
- Según la muestra tomada, el Segmento de CTéc hace un desglose por el campo Revistas, porque tiene diferentes Tipos de Documentos (TD).
- Los Tipos de Documentos con mayor ocurrencia dentro de la base de datos son las publicaciones Seriadas.

Otro de los resultados obtenidos fue tomando la muestra de los 10 492 registros y se ejecutó un nuevo árbol de decisión con los parámetros de id-column y label-column y los valores de las variables tomadas, Segmento y Keyword1 respectivamente, (anexo 14).

Los patrones de comportamiento que brinda este árbol son:

- El Segmento Csoci muestra que la etiqueta de mayor ocurrencia es Cuba, aunque indica por la coloración de la hoja la presencia de otros metadatos. La palabra 'Cuba' como índice de mayor ocurrencia en los metadatos, es una asignación que está restando la posibilidad de saber más sobre el tema que se ha investigado en este Segmento.

- Los Segmentos Biomed y CAgrop, muestran las Keyword1 de mayor ocurrencia, como 'Métodos' y 'Variedades' respectivamente, pueden estar pasando por lo mismo que el Segmento Csoci, porque no se puede apreciar por esta vía, sobre las variedades y métodos que se ha investigado al ser términos muy generales. Se propone revisar la asignación de los metadatos en estas publicaciones.

- El Segmento CTéc tiene un desglose por Revistas, para mostrar ordenadamente las diferentes Keyword1 que tiene.

- Cada Revista a través de la información de la Keyword1, destaca sobre la temática que más ha publicado.

- El Segmento CTéc tiene al menos una ocurrencia que responden a la 'Espectrofotometría' y esta a su vez tiene el campo Revista vacío porque es un MD.

- Se hizo una revisión de la relación existente entre la Revista y los metadatos y se comprobó que la Revista Nucleus asociada a la Keyword1 el metadato 'Agricultura', se justifica porque esta revista aunque pertenece a la Agencia de Energía Nuclear y Tecnologías de Avanzada, tiene como objetivo divulgar la ciencia y tecnologías nucleares de Cuba y el mundo, con sus principales aplicaciones en salud, agricultura, industria y medio ambiente. Otro ejemplo es la Revista Tecnología Química asociada a la Keyword1 'Trichodermo Viride' (Hongo y biofungicida para el control de enfermedades de plantas). En ambos casos se comprueba que las revistas publican artículos que están dentro de sus temáticas.

Después de hacer una segunda limpieza en los datos y considerando que el algoritmo que se está ejecutando es un Árbol de Decisión que está dentro del grupo de las técnicas de aprendizaje supervisado, se decidió ejecutarlo nuevamente, con el mismo juego de variables, el Segmento y Keyword1 para el id-column y label-column respectivamente y se obtuvo un árbol diferente, (anexo 6).

En este nuevo árbol se puede apreciar que se ha agregado nueva información, como es en el Segmento de CTécn el desglose por Tipo de Documento (TD), información que demuestra la razón por la cual no tienen información dentro del campo Revistas, es decir, consiste en que es un manuscrito depositado (MD) y no una Seriada que si son publicadas en revistas, quedando así justificado este campo vacío.

Se realizó una tercera limpieza en los datos, para asegurar que no quedara ningún error en los mismos, porque se sabe que los errores dentro de la información, pueden influir directamente en los resultados del Árbol de Decisión, es decir pueden excluir de una u otra clasificación a estos registros con errores o también pueden ser agregados a grupos a los que no pertenecen y se ejecutó el mismo algoritmo del árbol, pero ahora utilizando el juego de variables de Segmento y Revistas para el id-column y label-column respectivamente.

El árbol permite analizar los patrones de comportamientos siguientes, (anexo 7).

- Segmento *Csoci*, continúa destacándose la Keyword1 'Cuba', y el Segmento *CAgro*, se mantienen asociado a la Keyword1 'Variedades', no permitiendo ambos casos distinguir los tipos de investigaciones que desarrollan estas publicaciones. En ninguno de estos dos Segmentos se ha logrado un desglose por Revistas.

- Segmento *Biomed*, si hace un desglose por revistas, donde la Revista Cubana de Obstetricia y Ginecología, se desglosa por Idioma y los metadatos de mayor ocurrencia en español (spa) 'Biofísica' y en español e inglés (spa. eng) 'Obstetricia', además existe un grupo que tiene vacío el campo Revista y su metadato es 'Cerebro'. Se comprueba dentro de la base de datos que esta publicación está correcta porque está asociada a un premio académico (PAC).

- Segmento *CTéc,* la revista Nucleus hace un desglose por Idioma y los metadatos de mayor ocurrencia en español (spa) 'Agricultura' y en español e inglés (spa. eng) 'Irradiación', es decir esta revista publica en 2 idiomas.

- La Revista Ciencia, Innovación y Desarrollo hace un desglose por Keyword1, destacando que el mayor número de ocurrencias han estado asociadas a los Cambios Climáticos relacionado al metadato 'Medio Ambiente' y también el Medio Ambiente relacionado al metadato 'Contaminación'.

- Se observa que la hoja que tiene el metadato 'Espectrofotometría', sigue mostrando el campo Revista vacío y está comprobado en la base de datos, que es un manuscrito depositado (MD).

Como se pudo apreciar en estos resultados, los patrones no son información que se encuentran en ninguno de los campos, los patrones son el comportamiento de la información en cada caso, bajo determinadas condiciones y el conocer ese comportamiento permite tomar acción sobre el futuro en dependencia a lo que se quiere.

Otro resultado alcanzado fue a través del Meta Data View, (anexo 2), y del Text View, (anexo 8), que consiste en tomar cualquier Segmento del Text View que se desea analizar y extraerle los valores de los metadatos mayores que cero (>0) para una tabla y con el total de cada Segmento que muestra Meta Data View, en este caso CAgrop (99), calcular el porcentaje (%) de representación de estos datos en cuanto al valor total. De esta forma se puede obtener una métrica, para conocer sobre las investigaciones más trabajadas, (tabla 4).

Metadato	Ocurrencias
PASTIZAL PERMANENTE	1
ENGORDE	4
CRECIMIENTO	4
RAZAS (ANIMALES)	2
ALIMENTACION DE LOS ANIMALES	1

Tabla 4. Metadatos con una y más ocurrencias en CAgrop (Ejemplo 1). Fuente: elaboración propia.

En el ejemplo mostrado en la tabla 4, se observa que existen cuatro investigaciones sobre 'Engorde' y cuatro investigaciones sobre 'Crecimiento', este número es mayor comparado a las investigaciones realizadas sobre 'Pastizal Permanente' y ' Alimentación de los Animales', que tiene una investigación por cada una. Bajo un simple análisis se puede plantear que se investigó más sobre el 'Engorde' y el 'Crecimiento de los animales', que por la 'Alimentación de los animales'. Esta relación lógica de metadatos como "Alimentación, Crecimiento y Engorde" podría revisarse por los investigadores interesados, para asegurar que con la alimentación ya está todo estudiado, porque puede ser que la investigación sobre 'Alimentación de los animales' no sea para animales que se alimenten de 'Pastizales'.

La mayoría de los países que se destacan en materia de Ciencia y Tecnología, se basan en estos tipos de métrica para reasignar fondos para estudios e investigaciones, por ejemplo, los EE.UU a través de la ley del Congreso *"Goverment Performance And Results Act"* de 1993, evalúan de forma sistematizada la actividad científica, para poder rendir cuentas sobre el financiamiento otorgado por los programas de gasto público en investigaciones. Esta evaluación se hace a través de informes que recogen principalmente indicadores que justifican el uso del financiamiento por este concepto (USA, 2018). Dentro de estos indicadores se encuentran dos que vale destacar en esta investigación:

Los 'Indicadores De Producción y Producción Científica'. Este indicador sirve para medir la producción científica de un país o región. Según criterios de los especialistas, este indicador se obtiene contando el número de publicaciones, que dan cuenta de los resultados de las investigaciones hecha por un país o una región en particular (Guisán et al., 2006) .

Los 'Indicadores de Especialización Científica' recogen la distribución de las publicaciones de los países por campo científico y sirven para medir el peso de cada campo en un país, en comparación con su peso medio en el mundo.

Algunos estudiosos de estos indicadores plantean que las notables diferencias entre los distintos países, se justifica debido a las características de sus respectivas políticas científicas, que apoyan de manera diferente los distintas áreas científicas.

En el caso de los países en desarrollo los programas de investigaciones, están muy comprometidos con el desarrollo nacional, y de ahí que estas evaluaciones sirvan como instrumento para establecer prioridades en la asignación de recursos a centros de investigación o universidades, esto se corrobora con los países desarrollados como USA, que integra estas evaluaciones en proyectos (Erbschloe, 2017) y planes económicos (Erbschloe, 2018).

En esta investigación los patrones encontrados en cada Segmento, sirvieron para corroborar la necesidad de un estudio sobre estos indicadores. Estos patrones son producto de la clusterización de los metadatos asociados a las distintas publicaciones o investigaciones y su número de ocurrencias permiten conocer el tipo de actividad científica que han venido desarrollando los autores con mayor o menor interés, en un periodo de tiempo.

Se tomó otro grupo de metadatos de CAgrop, para mostrar que también la alimentación de los animales herbívoros, tiene menos investigaciones científicas, comparado con el 'Engorde' y 'Crecimiento' de las tablas 4 y 5.

Metadato	Ocurrencias
GANADO BOVINO	1
DIGESTION RUMINAL	2
PASTOREO ROTACIONAL	1
PASTOREO	1
PASTIZAL NATURAL	1

Tabla 5. Metadatos con una y más ocurrencias en CAgrop (ejemplo 2).
Fuente: elaboración propia.

En este ejemplo 2 habría que analizar si el ‘Engorde’ se refiere al ‘Ganado Bovino’ o a cual tipo de ganado, y sería interesante conocer la cantidad de investigaciones

nacionales que están relacionadas como para apoyar una investigación que abarque desde los diferentes pastizales por tipo de animales, sus digestiones, pastoreos, crecimiento y engorde, hasta la etapa adulta del animal, en el menor periodo de tiempo.

Todos estos resultados permiten conocer que se necesita investigar y se logró conocer con la ejecución de los árboles de decisión en el Rapid Miner V 4.6. También se tuvo en cuenta el algoritmo del árbol CHAID, para saber si pudiera existir diferencias en los resultados, pero como lo explica el propio Rapid Miner V 4.6, la diferencia entre estos dos árboles radica solo en que el árbol CHAID utiliza el Chi Cuadrado en su cálculo, mientras que el árbol de Decisión utiliza en el algoritmo el criterio de 'gain ratio', pero esto no cambia su resultado, (anexo 9).

Otra de las posibilidades del Rapid Miner V 4.6 que se utilizó fue el Plot View de la pestaña Data Table, ya que este permite obtener la gráfica del Self Organizing Maps (SOM), (anexo 4). Los datos de entrada entregados al Rapid Miner V 4.6 para su cálculo, están en el panel izquierdo de la gráfica y el algoritmo interno que utiliza para lograr estos resultados están basados en el "análisis de los K vecinos más cercanos" por Segmento. Como se observa, la propia herramienta distingue los cuatro Segmentos a los cuales da cuatro colores para diferenciarlos. Las dimensiones de red que utiliza para graficar esta herramienta son de Kohonen.net y el cálculo base lo realiza aplicando la 'distancia euclidiana o los K vecinos más cercanos', según explica la descripción que brinda el propio Rapid Miner V 4.6.

Esta gráfica permite interactuar con ella, ya que al pasar el puntero del mouse por encima de cada celda, muestra el nombre del Autor que representa. Es decir cada Autor está representado en una celda y la relaciona por el color al Segmento para el cual ha investigado. El color de cada uno de los Segmentos permite apreciar celdas que pertenecen a ese Segmento, como las que no pertenecen por tener colores diferentes.

De esta forma, rápidamente se puede conocer el nombre del Autor que se encuentra fuera de su Segmento y cuál es el Segmento en que debía estar ubicado.

Así, al tomar el nombre de un Autor que está fuera de su Segmento en el (SOM) y con una búsqueda de ese mismo Autor en el SGB, se obtiene un resultado que

comprueba a través de los metadatos los motivos por el cual ese Autor está fuera del Segmento. Esta validación cruzada de ambos resultados que pertenecen a herramientas con formas totalmente diferentes de calcular para lograr sus resultados, permitió comprobar en el SGB que ese Autor tiene publicaciones en más de un Segmento, por lo tanto ambas herramientas, sistema gestor de la biblioteca y la minería de datos, están trabajando correctamente, (anexo 4).

Esta primera demostración se realizó con la muestra completa de los 10 492 registros que tiene la base de datos, pero al igual que en el caso de los árboles se limitó la salida gráfica a 1000 casos solamente, por problemas de requerimientos tecnológicos de la computadora.

Los resultados que se presentan a continuación se seleccionaron al considerar que eran celdas que se encontraban en zonas de coloración diferente y a partir de esta condición la celda escogida fue al azar. La primera en el estudio representa al autor *Machado Noa; Noyla*. Con este nombre se hizo una búsqueda en el SGB y de esta búsqueda se recuperaron 3 publicaciones, dos de ellas responden a *Machado Noa; Noyla*, pero una no coincide porque pertenece a *Maceo Tames; F*, (anexo 15).

Se comprobó que verdaderamente es una mala recuperación en el SGB al mostrar al Autor Maceo Tames, F en la búsqueda de Machado Noa y además se revisó la base de datos directamente y se encontró que el Autor Machado Noa tiene 2 publicaciones en segmentos diferentes:

- Por Ciencias Agropecuarias - metadatos (Industria Azucarera, Gestión, Control de la Calidad, Administración de Empresas)
- Por Ciencias Sociales - metadatos (Gestión, Control de Sistema, Bancos, Cuba, Instituciones Financieras)

Es decir, el software Rapid Miner V 4.6 mostró al Autor *Machado Noa; Noyla,* correctamente dentro de los K prototipos que crea la red de Kohonen, que no es más que otra forma de agrupamiento.

Como se explicó anteriormente estos SOM, son un tipo de red neuronal no supervisada, donde el programa toma en cuenta los vectores de los datos introducidos para representar cada neurona y gana la que presenta menor diferencia, entre su vector de peso y el vector de datos. El cálculo de la distancia euclidiana

permitió comprobar la similaridad de estas dos publicaciones que están asentadas en segmentos diferentes de la ciencia, aunque pertenecen al mismo autor.

En relación a la recuperación errónea del SGB que se ha detectado por esta vía, se toma nota y se informa al Departamento de Bases de Datos, que es el encargado del mantenimiento y actualización del SGB, para que se repare este error en el menor tiempo posible.

Otra validación cruzada que se realizó fue con el nombre del Autor Melchor Orta; Gleiby. Para este caso se comprobó contra la base de datos, que este autor tiene 4 publicaciones y que todas están asentadas en el Segmento de las Ciencias Agropecuarias, por eso el color azul en su neurona, (anexo 16).
Las cuatro publicaciones que entregó el SGB fueron:

- Tesis (T)

  Metadatos: Rhizophora mangle, sanidad animal, Desinfectantes, Extracto vegetales
- Seriada (S)

  Metadatos: Rata, Toxicidad Aguda, Ligninas
- Seriada (S)

  Metadatos: Rhizophora mangle, Medicamentos Tradicionales, Temperatura, Taninos
- Seriada (S)

  Metadatos: Conejo (oryctolagus), sarna, enfermedades de la piel, miasis, dermotomicosis, terapéutica medicamentosa

El resultado brindado por el software Rapid Miner V 4.6 es un análisis correcto y la revisión directa de este Autor en la base de datos lo confirma, sin embargo cuando se hace una búsqueda de información en el SGB para este Autor, la recuperación es de un solo registro. Por lo tanto se toma nota, para verificar si realmente el SGB, está recuperando información de una de las versiones viejas de la base de datos, en la cual es posible que todavía no se encuentren incorporadas todas las publicaciones de este Autor, o si realmente está teniendo problemas en la recuperación y por lo tanto necesita mantenimiento. De comprobarse que el Sistema Gestor Bibliotecario

(SGB) tiene problemas, se debe revisar por los especialistas a cargo, para su reparación.

Con esto se demuestra que la validación cruzada es otra forma de detectar errores en el SGB, es una forma muy útil para detectar la asignación incorrecta de los metadatos en las publicaciones.

Otro tipo de resultado obtenido fue el gráfico de Distribución por Tipo de Documento (TD) en cada Segmento. El resultado de esta gráfica de barra, muestra la 'density/segmento', es decir densidad en el eje vertical y segmento en el eje horizontal, este tipo de gráfica es muy fácil de interpretar, (anexo 17).

Los patrones de comportamiento en las publicaciones según esta gráfica son:

CAgrop - no hay publicaciones por evento (E).

Biomed - no hay publicaciones por evento (E), ni manuscritos depositados (MD).

CTécn - no hay publicaciones por evento (E).

Csoci - hay de los cuatro tipos de documentos que recoge la base de datos.

Las causas que provocaron este tipo de comportamiento en los Segmentos, como es el caso del Segmento de Biomed, sin publicaciones de eventos E, ni manuscritos depositados MD o el caso de CAgrop y CTécn sin eventos E, se debe a que los especialistas de la Agricultura, Biomed y CTécn presentan sus trabajos en Eventos fuera de la biblioteca y por tanto esos trabajos y premiaciones no están registrados en la entidad.

Todos los resultados logrados con Rapid Miner V 4.6, se obtuvieron con la información de los campos Autor, Segmento, Tipo de Documentos y Keyword1 y en un caso la Keyword 2. Las Keywords restantes no se analizaron, porque tienen una cantidad significativa de registros vacíos. Todos los resultados fueron guardados en el repositorio para mejorar el servicio de boletines (DSI) y para estudios posteriores que apoyen la Vigilancia e Inteligencia y la Prospectiva.

WinIDAMS13-SP

Esta herramienta se emplea para realizar diferentes análisis de minería de datos y estadística a grandes volúmenes de información. En esta investigación se utilizaron los resultados estadísticos de WinIDAMS13-SP, para comprobar la efectividad de la metodología presentada en la figura 3 del capítulo II, a través de las mismas variables que se estudiaron con ambos softwares.

El WinIDAMS13-SP presenta la clusterización a través de las tablas multidimensionales, las estadísticas y además muestra diversas gráficas asociadas a estos resultados. La definición de la tabla multidimensional, que es el algoritmo base de donde se obtienen todos los otros resultados de WinIDAMS13-SP, está sujeta a una o más Variables de Página, Variables de Columna, Variables de Fila y Variables de Celda y este juego de variables o campos puede variar, tantas veces sea de interés para la investigación. Es decir con una misma base de datos se puede combinar diferentes juegos de variables para obtener diferentes resultados.

Para estos análisis se seleccionaron las mismas variables que se utilizaron con RapidMiner V 4.6 y se encontraron patrones interesantes, porque tanto las tablas multidimensionales como los arboles de decisión o la gráfica SOM utilizan la clusterización, que es la primera Tarea de Modelación que se planteó en la tabla 1 del capítulo I.

Las salidas de estas Tablas Multidimensionales son el resultado de la agrupación de la información según las Variables escogidas en celdas y adicionalmente calcula otros parámetros estadísticos tales como frecuencia, media y otros. Entre los resultados que se obtienen con esta herramienta están las gráficas de barras, los histogramas, que muestran en ventana emergente la regresión y la correlación entre variables (Ruiz et al., 2017).

Debe señalarse que para la mayoría de los cálculos a esta herramienta se le importó la base de datos completa con 10 492 registros, de los que aceptó 10 491 (anexo 19). En esta herramienta las muestras fueron tomadas al azar, una muestra con 8550 registros, otra con 3200 registros y otra con 2289 registros. No se calcularon los

tamaños de las muestras, porque se quiso investigar sobre la influencia del tamaño de las muestras sobre los resultados.

En el anexo 18, se presentan los resultados obtenidos a partir de la definición de Tabla Multidimensional, con las variables seleccionadas a ser procesadas:

Variables de Página: TD
Variables de Columna: Year (Año)
Variables de Fila: Segmento
Variables de Celda: Revista

En el anexo 19, aparece el 'Resumen Total de las Revistas por Segmento y Año', para los 10 491 registros. Esta tabla resultante tiene en la 1ra fila un encabezado por Year (Año), en la columna de referencia de la izquierda se observa un desglose de los cuatro segmentos, donde se muestran la 'Frecuencia' y la 'Revista media' y en la parte inferior y derecha se encuentran los totales de estas variables. Todas estas columnas y celdas se encuentran resaltadas por colores y los Totales están al final en azul, todo esto para hacer una distinción visual rápida. En la parte inferior de esta ventana hay un desglose por TD, es decir se puede acceder a siete ventanas, una por cada tipo de documento.

En el anexo 20, 'Resumen en Serie de las Revistas por Segmento y Year (Año)', se utilizó la base de datos que solo tiene 8 550 registros. Esta variación de registros en la base de datos, sirvió para comprobar que los resultados no están influenciados por el tamaño de la muestra. Ambos datos se pueden comprobar en la cifra que se encuentra en la celda última de la derecha, que muestra la suma de los Totales. También se aprecian los valores de Frecuencia y Revista-Media de cada Segmento por Year (Año).

Aunque los resultados de las tablas se pueden entender muy fácilmente, el software da la posibilidad de graficar estos resultados en un histograma. La preparación de un histograma se puede ver en el anexo 21, que escogió una salida con un estilo de 'Barra vertical 3D', para una prueba con 10491 registros. La gráfica está perfectamente identificada por cuatro colores, uno para cada Segmento, sobre un eje

horizontal dividido por años y un eje vertical dividido por cantidad de registros, como máxima salida toma hasta 1000 registros.

Un análisis de esta gráfica permite apreciar que existe un número grande de registros que tiene el campo Year (Año) en cero. Esta gráfica de barras permite analizar el comportamiento de las publicaciones por Year (Año) y Segmento, se puede ver que en el periodo entre 1999 al 2004 hubo un incremento en los cuatro tipos de publicaciones en la base de datos.

El análisis sobre el incremento de publicaciones entre 1999 - 2004, se justifica porque a partir del 1999, Cuba ingresa a la ALADI (Asociación Latinoamericana de Integración), (ALADI, 2015) y es que empieza a cambiar el panorama económico, comienzan a mejorar todo el país y con ello la comunidad científica, en relación a la investigación y la creación, anexo 21.

Un grupo de tablas multidimensionales se pueden ver en los anexos (22, 23 y 24)**,** donde en el (anexo 22) se define:

Variables de Página: Segmento
Variables de Columna: Year (Año)
Variables de Fila: Revista
Variables de Celda: TD

En el anexo 23 se obtienen todas las tablas y se muestra la Tabla Total, mientras que en el anexo 24 se muestra solo la Tabla de CAgrop. Este grupo de tablas muestran en el marco izquierdo un desglose de Frecuencia por Revista y la Media del TD.

En el anexo 25, el juego de variables fue:
Variables de Página: Segmento
Variables de Columna: Year (Año)
Variables de Fila: TD
Variables de Celda: Segmento

La tabla que se muestra es la tabla de Biomed y en el marco izquierdo se muestra un desglose por Tipo de documento (TD).

En el anexo 26 las variables estudiadas con 10492 registros fueron:

Variables de Página: Year (Año)

Variables de Columna: Segmento

Variables de Fila: Revista

Variables de Celda: Idioma

El anexo 26 muestra la página resumen de la estadística utilizada por el software durante la creación de las tablas multidimensionales. Estos valores estadísticos están agrupados por:

1. Ji-Cuadrado.
2. Medidas de asociación basadas en Ji-Cuadrado para variables nominales.
3. Medidas de asociación de variables ordinales.
   3.1 Medidas basadas en pares concordantes y discordantes.
   3.2 Medidas basadas en reducción proporcional en error.

La interpretación de este resultado del anexo 26 permite conocer el grado de relación entre dos o más variables cualitativas. En este caso se observa en el margen superior que el par de variables seleccionadas para el análisis estadístico fueron:

Fila: Revista y Col: Segmento.

1. El Ji-Cuadrado.

   Grados de libertad: 2968

   Ji-cuadrado:36059.74

   N-ajustado: 10492

El Ji-Cuadrado o Chi-Cuadrado, es nombrado por su autor como el coeficiente de correlación de Pearson, y se plantea que se emplea para “valorar la bondad del ajuste de unos datos a una distribución de probabilidad conocida, y se ha establecido como el procedimiento de elección para el contraste de hipótesis. Esta prueba estadística se emplea en el análisis de dos o más grupos y de dos o más variables. Desde entonces, se ha convertido en una prueba muy aceptada y aplicable a múltiples usos, cuando se dispone de datos independientes de tipo nominal. Esta ofrece un test general sobre la existencia de diferencias entre las categorías que agrupan a los datos de la variable dependiente” (Hernández de la Rosa et al., 2017).

En otras palabras, la correlación es la forma numérica en la que la estadística evalúa la relación de dos o más variables, es decir, mide la dependencia de una variable con respecto a otra variable independiente.

Por tanto, es una medida estadística que cuantifica la dependencia **lineal** entre dos variables, por lo tanto, si se representan en un diagrama de dispersión los valores que toman dos variables, el **coeficiente de correlación lineal** señalará lo bien o lo mal que el conjunto de puntos representados se aproxima a una recta.

2. Las Medidas de asociación basadas en Ji-Cuadrado para variables nominales.
   No requiere ningún orden de categorías de filas y columnas.
   Los valores mostrados en los resultados del anexo 26 fueron:

   Coeficiente Fi: 1.85
   Coeficiente de Contingencia: 0.88
   V de Cramer: 0.50

El análisis de estos valores plantea que el coeficiente Fi, es una medida que indica la intensidad de la relación entre dos variables dicotómicas. Cuando su valor es “0”, indica independencia entre las 2 variables, mientras que el valor “1” indica la mayor correlación entre las 2 variables, (Universidad-Barcelona, 2005), en este caso tiene 1.85.

Coeficiente de contingencia: 0.88

El coeficiente de contingencia de Pearson, expresa la intensidad de la relación entre dos o más variables nominales, su valor va entre -1 a +1, y estos extremos indican correlación máxima, mientras que el cero representa que no hay correlación alguna. En este caso el valor de 0.88 lo que indica que es una relación fuerte, ya que está próximo a +1.

V de Cramer: 0.50

El V de Cramer es una medida de relación estadística basada en el Ji Cuadrado (χ^2) y es usado para conocer la asociación de las variables nominales. Siempre está acotado entre 0 y 1 (sea cual sea la dimensión de la tabla).(Ramírez, 2018). Se dice que:

- V de Cramer = 0: no hay relación entre X y Y
- V de Cramer = 1: hay una relación perfecta entre X y Y

- V de Cramer = 0,6: hay una correlación relativamente intensa entre X y Y

Este resultado de 0.50, se puede considerar como una correlación por igual entre variables.

3. Las Medidas de asociación de variables ordinales.

3.1. Medidas basadas en pares concordantes y discordantes

Coeficiente de Tau-b de Kendall: -1.#J

Coeficiente de Tau-c de Stuart: 0.49

Gamma: 0.63

D asimétrica de Sormer - Var dep en columnas: 0.48

D asimétrica de Sormer - Var dep en filas: 0.61

D simétrica de Sormer: 0.54

El Coeficiente de Tau-b de Kendall se utiliza en tabulación cruzada para medir la asociación entre dos variables ordinales, por lo tanto, el valor negativo indica que ambas variables disminuyen simultáneamente. Los valores estadísticos establecidos de estas medidas están entre -1 a +1. La interpretación de estas tres medidas, indican que existe una asociación perfecta entre las variables que conforman la tabla que se analizó (Minitab, 2018).

Los tres valores sobre la D asimétrica de Somers miden la fuerza y la dirección de la relación entre pares de variables. Los valores de la D asimétrica de Somers van desde -1 (todos los pares son discordantes) hasta 1 (todos los pares son concordantes). En este caso los resultados muestran que la variable de columna (Segmento e Idioma) y la de fila (Revista) son concordantes y que la de mayor concordancia es la fila, en este caso contiene la variable de las Revistas, anexo 26.

3.2 Las Medidas basadas en reducción proporcional en error

Lambda asimétrica - Var dep en filas: 0.01

Lambda asimétrica - Var dep en columnas: 0.82

Lambda simétrica: 0.38

Lambda es la medida en reducción proporcional en error, mide la mejora porcentual en la probabilidad de la variable dependiente dado el valor de otras variables. Los

valores de Lambda van desde 0 hasta 1. Un valor de 0 significa que la variable independiente no mejora la predicción de las categorías de la variable dependiente. Un valor de 1 significa que la variable independiente predice completamente las categorías de la variable dependiente. Un valor de 0.5 significa que el error de predicción se reduce en 50% (Minitab, 2018).

En este caso la variable independiente es la columna que corresponde con la variable Segmento y los valores resultantes de Lambda asimétrica mayor es el de la columna con 0.82, por tanto como se sabe el Segmento determina el comportamiento de la variable dependiente, en este caso la variable Revista, anexo 26.

Estos resultados confirman que la variable Revista dependen primeramente del Segmento y en segundo lugar del Idioma, mientras que la Variable de Página, en este caso Year (Año), es la variable por la cual se ordenan los datos en la página, mientras que la variable que rellena la celda es la variable TD. Es decir en este juego de variables analizado confirma, que la variable Revista depende fundamentalmente del Segmento y el Idioma, es un juego de datos que está correctamente seleccionado para estos estudios.

El anexo 27 presenta los resultados de la Exploración gráfica de los datos y las variables que se estudiaron fueron, Segmento, Revista y Autor. Este tipo de gráfica es interactiva, por cada punto que se señale con el mouse, emerge una ventana con información estadística adicional, como se observa en el cuadro de la esquina superior izquierda. Tanto por la variable horizontal, como por la variable vertical muestra la media, la desviación estándar y el número de valores faltantes, además de mostrar 3 tipos de líneas, uno para la regresión lineal, otra para el análisis de regresión lineal local y otra para la media local.

El anexo 28 muestra los histogramas y gráficas de dispersión de las tablas multidimensionales, aquí se agregó el análisis de regresión y de correlación, con el objetivo de conocer el grado de relación entre las variables Revista, Segmento y Autor. La información estadística que ofrece este anexo, permite conocer sobre los valores de la asimetría (Skew), la kurtosis (Kurt) y la desviación estándar (Std).
Con este estudio se pudo comprobar que el juego de variables escogidos fue correcto para analizar y extraer patrones útiles, para estudios posteriores.

Los resultados logrados con ambos softwares RapidMiner V 4.6 y WinIDAMS13-SP, han servido para demostrar primeramente, que ambas herramientas son buenas e intuitivas para realizar este tipo de investigación, porque con RapidMiner V 4.6 se ha podido extraer patrones, que han cumplido con todos los objetivos de esta investigación y con WinIDAMS13-SP también se han extraído patrones, además que ha demostrado estadísticamente que las variables escogidas para el estudio, son las correctas, porque guardan relaciones entre ellas y sirven para estudios métricos futuros.

El desarrollo de esta fase 2 de la metodología presentada, ha permitido la aplicación de la minería de datos y la obtención de los patrones que se esperaban, se ha realizado la validación cruzada de los patrones con el sistema gestor bibliotecario (SGB), y con estos resultados se han arreglado todos los errores encontrados en la información de la base de datos, a la vez que ha servido para mejorar los productos y servicios, que brinda la Biblioteca con su SGB. Se han logrado patrones que proporcionan diferente conocimiento y se encuentran listos para pasar a la fase 3.

3.3 Fase 3. Resultados obtenidos en el Proceso de almacenamiento

La fase 3 de la Metodología comienza con el “Almacenamiento de la Información” que aportó la minería de datos. Para lograr esto se creó un repositorio digital utilizando la herramienta Eprint. Este repositorio tiene el objetivo de proteger y publicar los patrones encontrados en la fase 2 de la metodología desarrollada, estos resultados son los primeros en guardarse y son imprescindibles, para potenciar los nuevos productos y servicios que se quieran implementar.

A partir de aquí este repositorio pasa a formar parte de las herramientas que apoyan el trabajo de la biblioteca y en especial del Observatorio del Instituto, ya que permite cumplir con los objetivos de automatizar las búsquedas de información, ahorrar tiempo, detectar tendencias en el entorno científico, identificar oportunidades, conseguir nuevos clientes, que son los objetivos, tanto de la Biblioteca, como los del Observatorio.

Para renovar constantemente el valor del repositorio se mantiene la vigilancia a las bases de datos especializadas y base de datos de patentes, con vista a detectar investigaciones similares a la información registrada o información sobre nuevas

ramas de las ciencias que puedan estar vinculadas y que sean solicitada por alguna institución, anexo 29.

Un ejemplo, que sea el metadato seleccionado “abejas”, el cual se buscó en el sitio de WIPO, que es una colección de Patentes Internacionales. Inicialmente la búsqueda se filtró para que la información recuperada perteneciera solo a Cuba, con vistas a conocer primeramente cuántas investigaciones patentadas tenían los especialistas cubanos. Esta vigilancia es útil porque permite la colaboración entre investigadores (Utrilla, 2018) y de este modo evitar pérdida de tiempo, estudiando lo que ya se ha investigado e inclusive se encuentra patentado por otras instituciones del país. El resultado de esta búsqueda mostró que existen 7 investigaciones cubanas sobre las abejas que están patentadas y que por supuesto pasaron a formar parte de la información del repositorio.

Esto demuestra que el repositorio creado en la fase 3 tiene información preparada para realizar trabajos de vigilancia, inteligencia y prospección. Con esta información se puede conocer el comportamiento del metadato seleccionado dentro y fuera del país. En este ejemplo se puede decir que son investigaciones cubanas del 2003 y 2005, que ya tienen algunos años y que además hay que revisar en el caso de autores extranjeros, si fuera de interés de algún investigador, para detectar tendencias en el entorno y buscar nuevas oportunidades de investigación o de negocio dentro y fuera del país.

Con este nuevo repositorio se logra la segunda novedad científica planteada para esta investigación y la biblioteca tiene una herramienta que le sirve para crear diferentes productos y servicios.

3.4 Asimilación del Conocimiento

Esta etapa es nombrada así por sus creadores (Cabena et al., 1998) y es aquí donde se asimila toda la información del repositorio y el conocimiento que se deriva de esta información se toma para la creación de los nuevos productos y con ellos ofrecer servicios de publicación de patrones, de boletines DSI para los suscriptores, la vigilancia e inteligencia y la prospectiva como lo indica la metodología.

El repositorio creado tiene el potencial para apoyar la vigilancia e inteligencia y la prospectiva, y aunque estos son proyectos del observatorio institucional, que están

fuera de esta investigación, necesitan contar con información totalmente actualizada para poder tomar decisiones y por tanto, la biblioteca es el mejor lugar para facilitarle esta herramienta.

El servicio de Boletines de Difusión Selectiva de la Información (DSI) que ofrece la biblioteca habitualmente es un servicio de correo con la información solicitada por los usuarios inscritos, pero con la nueva información del repositorio, se agrega la posibilidad de un nuevo servicio que puede ser un sistema de boletines periódicos, con el objetivo de publicar las investigaciones más recientes de cualquiera de los Segmentos.

Estos resultados se pueden ofrecer en diferentes formatos como bases de datos, multimedia o en libros electrónicos, este último, para que puedan ser distribuidos para los diferentes dispositivos móviles.

Este tipo de publicación es una forma de gestionar la nueva información, revitalizar los servicios de la biblioteca y atraer nuevos usuarios.

3.5 Monitoreo y Control

Este es el último proceso de la fase 3 y como se explicó en el capítulo II tiene tres objetivos:

- El primer objetivo consistió en definir la cantidad mínima de registros que se debe incorporar a la base de datos para comenzar un nuevo estudio. Para esto se utilizó la fórmula de Pickers (fórmula 1), y su resultado fue de 100 registros, como mínimo.

- El segundo objetivo consistió en aplicar una encuesta realizada al personal bibliotecario para el control y monitoreo sobre la satisfacción de la metodología aplicada. Esta encuesta constó de 5 preguntas y cada una de ellas estuvo dirigida a una etapa de la metodología desarrollada, el análisis de estos resultados se hizo utilizando el software Statistic Program for Social Sciences IBM SPSS Statisticsv.21**,** anexo 30.

El número de encuestados fue de 23 personas. Esta encuesta se trabajó en Google y los resultados de las encuestas se exportaron a una tabla de Microsoft Excel y se guardaron en un fichero .cvs, este fichero fue importando en el software estadístico

IBM SPSS Statistics v.21, donde se analizó **la** frecuencia de las variables y su porcentaje, anexo 30.

El resultado de cada una de estas variables que dio el software IBM SPSS Statistics v.21 fue de la siguiente manera:

1. 1 El trabajo de selección y preparación de los datos antes de la importación fue?

23 respuestas de 23 encuestados.

Sencillo = 39.1%, Complejo = 39.1%, Muy Complejo = 21.7%

V1 = Un número de (9) personas consideran la selección y preparación de los datos como un trabajo "Sencillo", mientras otro número igual lo consideran complejo.

2. La limpieza y estandarización de la información benefició a la BD?

18 respuestas de 21 encuestados Si = 85.7% Tal vez = 14.3%

V2 = El mayor número de personas (18) consideran como un beneficio la limpieza y estandarización realizada a la información.

3. La limpieza y estandarización de la información benefició al sistema gestor de la biblioteca.

19 respuestas de 23 encuestados Si = 82.6 % Tal vez = 4 %

V3 = El mayor número de personas (19) consideran como un beneficio la limpieza y estandarización realizada a la información del sistema gestor bibliotecario.

4. ¿Cree que el estudio con minería de datos mejoró los productos y servicios bibliotecarios?

19 respuestas de 22 encuestados Si = 86.4 % Tal vez 13.6 %

V4 = El mayor número de personas (19) consideran que el estudio de minería de datos **mejoró los productos y servicios bibliotecarios**.

5. ¿Cómo considera la metodología aplicada para el estudio de datos bibliográficos?

14 respuestas de 23 encuestados

Beneficiosa: 60.9 % Buena: 8 % Mala: 4.3 %

V5 = El mayor número de personas (14) consideran como Beneficiosa a la Metodología aplicada para el estudio de Datos Bibliográficos

El resumen de esta encuesta analizada con el software estadístico IBM SPSS Statistics v.21**,** confirma que se cumplieron los 4 beneficios internos, propuestos en la Introducción de este trabajo, de la siguiente manera, (tabla 6).

Beneficios Internos	Resultados de la encuesta
Mejorar la limpieza y estandarización de la información benefició a la BD	**El trabajo de selección y preparación de los datos, es considerado Sencillo = 85.7% por 18 personas de 21**
Mejorar el proceso de recuperación de la información en el sistema gestor de las bibliotecas	**El proceso de limpieza y estandarización beneficio al SGB, Si = 82.6% por 19 personas de 23**
Mejorar la oferta de productos y servicios que ofrece la biblioteca	**El estudio con MD mejora los productos y servicios, Si = 86.4% por 19 personas de 22**
Lograr una gestión eficiente de la información	**La Metodología aplicada al estudio, por resultado de la encuesta es, Beneficiosa = 60.9 % por 14 personas de 23**

Tabla 6. Encuesta al personal bibliotecario.
Fuente: elaboración propia.

Esta encuesta dirigida al personal de la biblioteca mostró que los 4 beneficios internos propuestos en la Introducción se lograron, además que demostró estadísticamente, que la Metodología aplicada es beneficiosa en un 60.9%.

- El tercer objetivo propuesto: consistió en aplicar una encuesta a los clientes externos, con el objetivo de conocer el nivel de satisfacción en relación a los servicios y productos que ofrece la biblioteca y para esto se utilizó una encuesta que fue basada en la Técnica de Iadov.

Esta técnica de Iadov constituye una vía indirecta para el estudio de la satisfacción, ya que los criterios que se utilizan se fundamentan en las relaciones que se establecen entre tres preguntas cerradas que se intercalan dentro de un cuestionario (en este estudio preguntas 4, 5 y 6) y cuya relación el encuestado desconoce. Estas

tres preguntas se relacionan a través de lo que se denomina el "Cuadro Lógico de Iadov", además existen otras 5 preguntas con las cuales se quiere conocer otras necesidades de los clientes, (anexo 31).

Las respuestas de las 30 encuestas se observan en el anexo 32, en el mismo se resaltan las tres preguntas relacionadas en color rojo, para facilitar la confección del Cuadro lógico de Iadov, (anexo 33).

La Técnica de Iadov plantea una fórmula que mide el Índice de satisfacción grupal (ISG), basado en la frecuencia de cada respuesta que es llevado a la tabla y de ahí a la formula, por lo tanto, el nivel de satisfacción de los clientes externos resultó ser un valor de 0.90.

Por lo tanto, esto confirma que la metodología para el estudio de datos bibliográficos con el empleo de la minería de datos, elevó la calidad de productos y servicios de la biblioteca, y a su vez proporcionó un mayor nivel de satisfacción del cliente externo.

Conclusiones parciales

1. Los árboles de decisión permitieron obtener los patrones de comportamiento de la información, además de los clústeres de información más representativos para estudios futuros.

2. Los Self Organizing Maps (SOM) permitieron conocer los Autores que se encuentran fuera del Segmento, y decidir con la validación cruzada, si era correcta o no los resultados del Sistema Gestor Bibliotecario.

3. Se pudo comprobar la veracidad de los resultados logrados con el software de minería de datos, *RAPID MINER V 4.6* y los resultados estadísticos de *WinIDAMS13-SP*, demostraron que fue escogida la mejor combinación de variables para la investigación.

4. La Encuesta realizada al personal bibliotecario, permitió comprobar que la Metodología aplicada es beneficiosa en un 60.9%.

5. La Técnica de Iadov permitió conocer que los nuevos productos y servicios logrados a través de la *metodología para el estudio de datos bibliográficos con el empleo de la minería de datos* reflejan un Índice de satisfacción grupal del 0.90.

Conclusiones

1. Se propone una metodología para el estudio de datos bibliográficos con el empleo de la minería de datos en la Biblioteca de Ciencia y Técnica para la gestión efectiva del conocimiento y sus servicios que tributan al cumplimiento de los objetivos definidos.

2. El procedimiento aplicado en la metodología se caracteriza por su capacidad de despliegue y herramientas que aportan la eliminación de los problemas que presentaba la información dentro de la base de datos y que ocasionaban errores al Sistema de Gestión Bibliotecario durante la recuperación de información.

3. La metodología propuesta es novedosa, se fundamenta desde la categoría conceptual de la minería de datos para una biblioteca en la que se define formas de realización, utilizando técnicas que facilitan la oportuna búsqueda en el fondo bibliotecario como parte del servicio de apoyo a los diferentes estudios de la ciencia, la tecnología, la innovación y el medio ambiente, solicitados a la Biblioteca de Ciencia y Técnica.

Recomendaciones

1. Dar continuidad a la investigación en el ámbito de procederes, métodos y herramientas para gestionar el conocimiento que permitan incrementar la pertinencia, robustez y viabilidad del procedimiento general propuesto y que permita el monitorio y control de la información, para el mejoramiento de la Biblioteca.

2. Continuar con la divulgación de los resultados obtenidos en la investigación a través de presentaciones en eventos científicos, artículos, libros, tesis y cursos de formación / capacitación con vistas a extender estos resultados, progresivamente a otras bibliotecas con sus correspondientes adecuaciones y ajustes.

Referencias bibliográficas

Acosta, C. M. C. (2019). Minería de Datos para Consolidación de Estados Financieros en Compañías de Outsourcing Contable. from http://biblioteca.uteg.edu.ec:8080/bitstream/handle/123456789/977/Miner%C3%ADa%20de%20datos%20para%20consolidaci%C3%B3n%20de%20estados%20financieros%20en%20compa%C3%B1%C3%ADas%20de%20Outsourcing%20contable..pdf?sequence=3&isAllowed=y

Aguilar, L., Victoria, Miguel Angel (2016). Sistema de razonamiento basado en casos, para la mejora de atención de salud en un centro rural Tesis: Facultad De Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú. from http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/5690/Aguilar_lvm.pdf?sequence=1

Al-Attar, A. (1997). CRITIKAL: client-server rule induction technology for industrial knowledge acquisition from large databases. Project ID: 22700, United Kingdom , IEEE.org. from https://ieeexplore.ieee.org/document/583703/authors#authors

ALADI. (2015). ALADI - Asociación Latinoamericana de Integración. Organización Mundial Del Comercio. from http://www.aladi.org/sitioaladi/quienessomos.html

Alcázar Román, J. M. (2023). EVALUACIÓN DEL ANÁLISIS DE CAPACIDAD DE UN CASO INDUSTRIAL PARA DATOS NO NORMALES. Logos. Estudio de caso, Repositorio Vol. 4 No. 2 ISSN 2215-5910. from https://dspace.ulead.ac.cr/repositorio/bitstream/handle/123456789/252/Jos%C3%A9Alc%C3%A1zar.pdf?sequence=1&isAllowed=y

Angulo, M. L. E. (2020). Redes bayesianas en R: análisis de los paquetes software disponibles from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi71dDfsMjzAhV2RTABHcCNCTsQFnoECAsQAQ&url=http%3A%2F%2Foa.upm.es%2F63644%2F1%2FTFM_LUIS_EDUARDO_ANGULO_MONTES.pdf&usg=AOvVaw0wz2cLGkCrNGfAGZmh5k0O

Arcos, M., P. , Mil, C., E. M. , López, G., L. A., & Pech, M., F. . (2019). Análisis de Minería de Datos y Machine Learning en Marketing Digital from https://www.academia.edu/41789925/An%C3%A1lisis_de_Miner%C3%ADa_de_Datos_y_Machine_Learning_en_Marketing_Digital

Arias, C. A. E. (2021). Modelo Interactivo De Visualización De Información Utilizando Librerías De Renderizado 3d En Aplicaciones Web, Aplicado A La Reducción De Dimensiones from http://repositorio.utn.edu.ec/bitstream/123456789/11523/2/04%20ISC%20595%20TRABAJO%20GRADO.pdf

Azevedo, A., & Santos, M. F. (2013). KDD, SEMMA AND CRISP-DM: A parallel overview. ISBN:978-972-8924-63-8. from https://pdfs.semanticscholar.org/7dfe/3bc6035da527deaa72007a27cef94047a7f9.pdf

Benitez, S., Ignacio Javier (2005). Tecnicas de Agrupamiento para el Analisis de Datos Cuantitativos y Cualitativos. Universidad Politecnica de Valencia. . from https://www.researchgate.net/profile/Ignacio-Benitez-3/publication/239526131_Tecnicas_de_Agrupamiento_para_el_Analisis_de_Datos_Cuantitativos_y_Cualitativos/links/00b7d51c15cca2cb1f000000/Tecnicas-de-Agrupamiento-para-el-Analisis-de-Datos-Cuantitativos-y-Cualitativos.pdf

Britez, L. (2021). Algoritmo para clasificación de vehículos mediante redes neuronales. from https://revistas.untref.edu.ar/index.php/innova/article/view/1133

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). Discovering Data Mining: From Concept to Implementation. Prentice Hall. from http://www.zanasi-alessandro.eu/publications/cabena-p-hadjinian-p-stadler-r-verhees-j-zanasi-a-1998-discovering-data-mining-from-concept-to-implementation/

Calvo Pérez, I. (2021). Algoritmos de aprendizaje automático para detección de fraudes con tarjetas de crédito: Análisis y comparativa. from https://oa.upm.es/67976/

Candás, R. J. (2006). Minería de datos en bibliotecas: bibliominería. from http://www.ub.edu/bid/17canda2.htm

Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. from https://onlinelibrary.wiley.com/doi/10.1002/bimj.202000196

Carpio-Martin, D. (2020). Mapas Autoorganizados (Self-Organizing Maps). from https://uvadoc.uva.es/bitstream/handle/10324/43792/TFG-G4605.pdf;jsessionid=38F8D0BCC8B6BB21822A627F58427871?sequence=1

Castillo, L. (2013). Tema 6.- Difusión de la información. from https://www.uv.es/macas/T6.pdf

CENATAV. (2023). Advanced Technologies Application Center. from http://www.cenatav.co.cu

CERPAMID. (2023). Centro de Estudios de Reconocimiento de Patrones y Minería de Datos (CERPAMID). from http://www.cerpamid.co.cu/index.php

Chagoya, R., Ena (2023). Métodos y técnicas de investigación. from https://www.gestiopolis.com/metodos-y-tecnicas-de-investigacion/

Chang, J., O'Reilly, C., Pontika, N., Owen, G., Haug, K., & Oudenhoven, M. (2018). ¿Qué es la minería de textos, cómo funciona y por qué es útil? , from https://universoabierto.org/2018/02/22/que-es-la-mineria-de-textos-como-funciona-y-por-que-es-util/

Chapman, A. (2012). Mineria de datos en Redes Sociales. La metodología PEIC, creada por Chapman. from http://mineriadatosredessociales.blogspot.com/2012/10/la-metodologia-peic-creada-por-chapman.html

Chen, J. (2021). Skewness. from https://www.investopedia.com/terms/s/skewness.asp

De la Puente, M. (2010). Bibliominería: bibliometría y minería de datos. *Consultora de Ciencias de la Información, Buenos Aires, Argentina.* from https://www.researchgate.net/publication/43139390_Bibliomineria_bibliometria_y_mineria_de_datos

De Volder, C. V. (2005). Los catálogos en línea de acceso público (OPACs) de las bibliotecas nacionales sudamericanas: evaluación y análisis comparativo. Instituto de Investigaciones Gino Germani, Facultad de Ciencias Sociales, UBA. from https://core.ac.uk/download/pdf/11884078.pdf

Díaz, M. J. d. J. (2021). Tema 1. Inteligencia de Negocio. . *Tema03-Modelos_de_Prediccion-Parte_II_Regresion-Y-Series-Temporales-2020-21.* from https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/Curso20-21/Tema03-Modelos_de_Prediccion-Parte_II_Regresion-Y-Series-Temporales-2020-21.pdf

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. from https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-021-00516-9.pdf

Erbschloe, M. (2017). Data Mining for National Security: U.S. Government Programs. free-ebooks.net. from https://www.free-ebooks.net/ebook/Data-Mining-for-National-Security-U-S-Government-Programs

Erbschloe, M. (2018). Big Data Technology In the U.S. Government free-ebooks.net. from https://www.free-ebooks.net/ebook/Big-Data-Technology-In-the-U-S-Government

Espinoza Hoyos, C. A. (2020). Simulación Y Evaluación De Técnicas De Clustering Para El Reconocimiento De Gestos Estáticos En La Traducción De La Lengua De Señas Peruana from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjEusaR87zzAhXqlHIEHSgvB1IQFnoECBAQAQ&url=https%3A%2F%2Frepositorio.urp.edu.pe%2Fbitstream%2Fhandle%2FURP%2F3647%2FELEC-T030_73892049_T%2520%2520%2520ESPINOZA%2520HOYOS%2520CARLOS%2520ANIBAL.pdf%3Fsequence%3D1%26isAllowed%3Dy&usg=AOvVaw2tU5zH8q51BIFVW6pRDbsQ

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine Volume 17 Number 3 (1996). from https://www.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131/0

Febles Rodríguez, J. P., & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. ACIMED 02 2002. from http://bvs.sld.cu/revistas/aci/vol10_2_02/aci03202.htm

Fernández, Lobelle, F., Gretel, & Rivera, Z. (2018). Las bibliotecas públicas por el desarrollo sostenible. Revista Cubana de Información en Ciencias de la Salud versión On-line ISSN 2307-2113. from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2307-21132018000200002

Fernández, E., Merlino, H., Ochoa, M., Diez, E., Britos, P., & García, M., R. . (2023). GESTIÓN ASISTIDA DE DOCUMENTOS EN UNA METODOLOGÍA DE EXPLOTACIÓN DE INFORMACIÓN. Facultad de Ingeniería. Universidad de Buenos Aires. from https://core.ac.uk/download/pdf/296349143.pdf

Flores, A. (2021). Técnicas y herramientas de minería de datos que cambiarán el rumbo de tu negocio. from https://www.crehana.com/blog/desarrollo-web/herramientas-mineria-datos/

Flores Rodrıguez, C. (2021). Generación de ́arboles de decisión usando un algoritmo inspirado en la Ísica. from https://www.uv.mx/personal/emezura/files/2021/03/Thesis-Camilo.pdf

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discoveryin Databases: An Overview. from https://pdfs.semanticscholar.org/7a7b/51b86e22d0077215287980c7ba793b09e4cd.pdf

Gallardo Arancibia, J. A. (2003). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. EPB 603 Sistemas del Conocimiento. ER-DM. from http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

García, S., Ramírez-Gallego, S., Luengo, J., & Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos. Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada. from https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf

Gea, M. M., Batanero, C., & Roa, R. (2014). El sentido de la correlación y regresión. Universidad de Granada. España. from https://www.researchgate.net/publication/282279255_El_sentido_de_la_correlacion_y_regresion

Gómez-Gil, M.-d.-P. (2021). Mapas Auto-Organizados. from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiuw6_9io30AhWRSjABHVgnB3kQFnoECAsQAQ&url=https%3A%2F%2Fccc.inaoep.mx%2F~pgomez%2Fcursos%2FIC-I%2Facetatos%2Fmapas.pdf&usg=AOvVaw0s21hxJXYT-ArhW9SezVHt

Gómez, F. A. J. (2014). Capítulo 2 : Estado del arte. 2.1 Introducción. from http://www.gsi.dit.upm.es/~anto/tesis/html/stateart.html ; http://dit.upm.es/~anto/tesis/html/stateart.html

González, A. (2021). Conceptos Básicos de Machine Learning. from https://cleverdata.io/conceptos-basicos-machine-learning/

González López, Manuel (2021). Pre-procesamiento de datos para aprendizaje de Distribución de Etiquetas. from file:///C:/Users/M/AppData/Local/Temp/88297.pdf

Graham Williams. (2010). Data Mining. Desktop Survival Guide. Copyright 2004-2010. Togaware Pty Ltd from http://datamining.togaware.com/survivor/index.html

Guisán, M. C., & Cancelo, M. T. (2006). INDICADORES DE PRODUCCIÓN CIENTÍFICA EN ECONOMÍA, CIENCIA Y TECNOLOGÍA: ANÁLISIS COMPARATIVO DE ESPAÑA, UNIÓN EUROPEA Y ESTADOS UNIDOS, 2001-2006 Estudios Económicos de Desarrollo Internacional, Vol. 6-2 (2006) from http://www.usc.gal/economet/reviews/eedi622pdf

Harmouch, M. (2021). 17 Clustering Algorithms Used In Data Science and Mining. from https://towardsdatascience.com/17-clustering-algorithms-used-in-data-science-mining-49dbfa5bf69a

Hernández de la Rosa, Y., Hernández, M., V. J., Batista, H., N. E., & Tejeda, C., E. . (2017). ¿Chi cuadrado o Ji cuadrado? E-ISSN: 10293043 | RNPS 1820. from http://scielo.sld.cu/pdf/mdc/v21n4/mdc01417.pdf

Hernández, S. R., Fernández, C. C., & Baptista, L. M. d. P. (2014). Metodología de la Investigación. 6ta Edicion. McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V., from http://observatorio.epacartagena.gov.co/wp-content/uploads/2017/08/metodologia-de-la-investigacion-sexta-edicion.compressed.pdf

Herrera, A. R. J., & Fontalvo, H. T. J. (2011). Seis Sigma Métodos Estadísticos y Sus Aplicaciones. from http://biblioteca.utec.edu.sv/siab/virtual/elibros_internet/55821.pdf

Herrera Varela, R. (2006). Bibliomining: minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. from http://bibliotecarios.cl/conferencia_2006/C2006_019.pdf

Herrero, C., Juan Carlos (2023). Comparativa Económica de las Comunidades Autónomas españolas en el siglo XXI: Un análisis multivariante. Universidad de Valladolid. from https://uvadoc.uva.es/bitstream/handle/10324/59600/TFG-E-1715.pdf?sequence=1&isAllowed=y

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2002). Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. Springer Link. from https://link.springer.com/chapter/10.1007/3-540-46131-0_2

Husnain, G., & Anwar, S. (2021). An intelligent cluster optimization algorithm based on Whale Optimization Algorithm for VANETs (WOACNET). from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250271

Ibarra, M. d. l. A. (2021). Tratamiento de los datos: OLTP, OLAP, Data Warehouse. *División consultoría de EvaluandoSoftware.com.* from https://www.evaluandosoftware.com/tratamiento-los-datos-oltp-olap-data-warehouse/

IBM. (2021a). Diagramas de dispersión y gráficos de puntos. SPSS Statistics. from https://www.ibm.com/docs/es/spss-statistics/beta?topic=types-scatter-plots-dot-plots

IBM. (2021b). Modelos de árboles de decisión. from https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=trees-decision-tree-models

IBM. (2023a). Dendrograma. SPSS Statistics. from https://www.ibm.com/docs/es/spss-statistics/29.0.0?topic=automobiles-dendrogram

IBM. (2023b). ¿Qué son las redes neuronales? . Soluciones IBM Cloud. from https://www.ibm.com/mx-es/topics/neural-networks

Jeré, A. (2018). Data quality in data warehouses. Lahti University of Applied Sciences Degree Programme in Information and Communications Technology. from https://www.theseus.fi/bitstream/handle/10024/146311/Aunola_Jere.pdf?sequence=2&isAllowed=y

KDnuggets, P. (2007). Data Mining Methodology (Aug.2007). from http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

KDnuggets Polls. (2019). KDnuggets Analytics/Data Science 2019 Software Poll. from https://www.kdnuggets.com/2019/05/new-poll-software-analytics-data-science-machine-learning.html

Kumar-Pandey, K., & Shukla, D. (2021). Euclidean distance stratified random sampling based clustering model for big data mining. from https://onlinelibrary.wiley.com/doi/10.1002/cmm4.1206?af=R

Lara Turrent, A. (2012). CAPITULO 4. METODOLOGIA. 4.1 METODOLOGÍA SIX SIGMA. from http://catarina.udlap.mx/u_dl_a/tales/documentos/lii/lara_t_a/capitulo4.pdf

León, E., Proaño, E., Muirragui, V., & Cajamarca, J. (2019). Minería de datos en el análisis de tendencias políticas en redes sociales. Ciencia digital, Vol. 3, N°3.4, p. 91 - 103, septiembre, 2019. from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjqs_aZ1KaAAxWsmYQIHWE9BI8QFnoECA0QAQ&url=https%3A%2F%2Fcienciadigital.org%2Frevistacienciadigital2%2Findex.php%2FCienciaDigital%2Farticle%2Fdownload%2F837%2F2023%2F&usg=AOvVaw0PJmCl3aFYhafae731pCj-&opi=89978449

Lumbreras, H., Maria Isabel. (2020). Evaluación de análisis de clustering jerárquico en datos moleculares de alta dimensión. Universidad de Cataluña. from https://openaccess.uoc.edu/bitstream/10609/120648/6/milumbrerasTFM0620memoria.pdf

Mantilla, A., Meneses, A., & Zúñiga, L. (2021). Estimación de la función de densidad no paramétrica para caracterizar los tiempos de reconfiguración empleados por los controladores de un prototipo de Robot Planar. from http://perspectivas.espoch.edu.ec:8081/index.php/RCP_ESPOCH/article/view/138/104

Martínez Heras, J. (2021). Árboles de Decisión con ejemplos en Python. from https://www.iartificial.net/arboles-de-decision-con-ejemplos-en-python/

Martínez, P. J. R., Ferrás, F. Y., Bermudez, C. L. L., Ortíz, C. Y., & Pérez, L. E. H. (2020). Regresión logística y predicción del bajo rendimiento académico de estudiantes en la carrera Medicina from http://www.revzoilomarinello.sld.cu/index.php/zmv/article/view/2230/pdf_691

Marulanda, E. C. E., López, T. M., & Mejía, S. M. H. (2017). Minería de datos en gestión del conocimiento de pymes de Colombia*. Revista Virtual Universidad Católica del Norte, 50, 224-237. from http://www.redalyc.org/pdf/1942/194250865013.pdf

Mateos, M. C. (2021). Definición y estudios de redes bayesianas aplicadas a ciencias de la salud y de la vida from https://eprints.ucm.es/id/eprint/68123/

Matos, G., Chalmeta, R., & Coltell, O. (2006). *Metodología para la Extracción del Conocimiento Empresarial a partir de los Datos*. Información Tecnológica. 17, No.2, from http://www.scielo.cl/scielo.php?pid=S0718-07642006000200011&script=sci_arttext

Maydana, H. A. R. (2021). Elección Del Mejor Modelo Entre Regresión Lineal Múltiple Y Árboles De Regresión Para Predecir El Precio Máximo De Las Acciones De Intel En Función Al Precio De Apertura Y Volumen De Ventas De Acciones Por Dia - 2019. from http://repositorio.unap.edu.pe/handle/UNAP/15333

Medina, V., Javier Enrique,. (2015). Los Estudios del Futuro y la Prospectiva: Claves para la construccion social de las regiones. Documento 96/32. Serie Ensayos. from https://repositorio.cepal.org/bitstream/handle/11362/9713/S9600704_es.pdf?sequence=1&isAllowed=y

Merkle. (2021). El algoritmo K-NN y su importancia en el modelado de datos. from https://www.merkle.com/es/es/blog/algoritmo-knn-modelado-datos

Merlino, H. (2004). Metodología de transformación de datos para su explotación. from http://laboratorios.fi.uba.ar/lsi/R-rtis-6-2-2004.pdf

Microsoft. (2014). Prueba y validación (minería de datos). *SQL Server 2014.* from https://msdn.microsoft.com/es-es/library/ms174493(v=sql.120).aspx

Microsoft. (2021). Gráfico de dispersión (Analysis Services - Minería de datos). from https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/scatter-plot-analysis-services-data-mining?view=sql-server-2017

Michalski, R. S. (1986). Concept Learning. from www.mli.gmu.edu/papers/86-90/86-17.pdf

Migriño, J. R., & Batangan, A. R. U. (2021). Using machine learning to create a decision tree model to predict outcomes of COVID-19 cases in the Philippines. from https://ojs.wpro.who.int/ojs/index.php/wpsar/article/view/831/1050

Minitab. (2018). Medidas de asociación para Tabulación cruzada y chi-cuadrada. Copyright © 2023 Minitab, LLC. All rights Reserved., from https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistics/tables/how-to/cross-tabulation-and-chi-square/interpret-the-results/all-statistics-and-graphs/measures-of-association/

MINJUS. (2010). GACETA OFICIAL DE LA REPÚBLICA DE CUBA, MINISTERIO DE JUSTICIA Gaceta Oficial No. 030 Ordinaria de 10 de agosto de 2010. Universidad Virtual de Salud de la Facultad de Ciencias Médicas Manuel Fajardo. from http://uvsfajardo.sld.cu/decreto-ley-271-de-bibliotecas

MINJUS. (2021). Decreto-Ley 7/2020 "Del Sistema de Ciencia, Tecnología e Innovación" (GOC-2021-765-O93). Gaceta Oficial No. 93 Ordinaria de 18 de agosto de 2021. from https://www.3ce.cu/sites/default/files/2023-01/decreto-ley-7-2020-del-sistema-de-ciencia-tecnologia-e-innovacion.pdf

Moine, J. M. (2017). Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. from https://core.ac.uk/download/pdf/16703288.pdf

Nicholson, S. (2003). Avoiding the Great Data-Wipe of Ought-Three: Maintaining an Institutional Record for Library Decision-Making in Threatening Times. from http://bibliomining.com/nicholson/preshred.htm

Nicholson, S. (2003 b). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. Information Technology and Libraries 22 (4), (tentative preprint) from http://bibliomining.com/nicholson/biblioprocess.htm

Nicholson, S. (2006). The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services. from http://scottnicholson.com/pubs/nicholsonbibliointro.pdf

Nicholson, S., & Stanton, J. (2003). Gaining strategic advantage through bibliomining: data mining for management decisions in corporate, special, digital, and traditional libraries. from https://surface.syr.edu/istpub/102/

Nieto, J. A. (2021). Algoritmos de Aprendisaje Automatico. Un Estudio de Difusion y Utilización. from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiY47-6-630AhWLSzABHbpgAb0QFnoECCEQAQ&url=https%3A%2F%2Foa.upm.es%2F68484%2F1%2FTFG_ALEJANDRO_NIETO_JEUX.pdf&usg=AOvVaw1_kZoMgAzc9ZwCNLZofwCF

Pascual, D., Pla, F., & Salvador, S. J. (2006). Hierarchical-Based Clustering Using Local Density Information For Overlapping Distributions. from http://marmota.dlsi.uji.es/WebBIB/papers/2006/Pascual-2006-RedRFbook.pdf

Peña, P. G. M. (2015). Diseño de una Arquitectura de Inteligencia de Negocios para e Área de Compras de Seguros Bolívar. UNIVERSIDAD LIBRE DE COLOMBIA. from https://repository.unilibre.edu.co/bitstream/handle/10901/8914/dise%C3%B1o%20arquitectura%20de%20negocios.pdf?sequence=1

Pickers, S. (2015). ¿Cómo determinar el tamaño de una muestra? Copyright © 2018 PSYMA GROUP AG, Rueckersdorf / Nuremberg from http://www.psyma.com/company/news/message/como-determinar-el-tamano-de-una-muestra

Pio, A. S. (2018). MINERÍA DE CALIDAD DE DATOS: APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA EVALUACIÓN DE LA CALIDAD DE LOS DATOS. UNIVERSIDAD DE LA REPÚBLICA. MONTEVIDEO – URUGUAY. from https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/25468/1/PIO18.pdf

Ramírez, M., Arturo (2018). 3.6 COEFICIENTE V DE CRAMER - Tecnicas de Investigacion Educactiva G38. from https://sites.google.com/site/tecnicasdeinvestigaciond38/estadisticas-no-parametricas/3-6-coeficiente-v-de-cramer

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & DeBastiani, F. (2019). Skewness and kurtosis comparisons of continuous distributions. from https://www.researchgate.net/publication/340157930_Skewness_and_kurtosis_comparisons_of_continuous_distributions

Rojas, P. R. J. (2021). Modelo de Aprendizaje Automático Supervisado para Identificar Patrones de Bajo Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico

Público – Juliaca from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwisxtDJkPnzAhXaQzABHXHvBgIQFnoECBgQAQ&url=https%3A%2F%2Frepositorio.upeu.edu.pe%2Fbitstream%2Fhandle%2F20.500.12840%2F4505%2FRudy_Tesis_Licenciatura_2021.pdf%3Fsequence%3D5%26isAllowed%3Dy&usg=AOvVaw0TqbJ3GeOalHboF-P4sIYD

Roman, V. (2019). Aprendizaje No Supervisado en Machine Learning: Agrupación. from https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

Romero, J. (2019). Técnicas y algoritmos de Minería de Datos. from https://jorgeromero.net/tecnicas-y-algoritmos-de-mineria-de-datos/

Romero, J. (2019a). Análisis Clúster. from https://jorgeromero.net/analisis-cluster/

Rouse, M. (2021). Visualización de datos. from https://www.computerweekly.com/es/definicion/Visualizacion-de-datos

Roy, G., Ivonne, Rivas, R., Rodolfo, Pérez, R., Marcela, & Palacios, C., Lino. (2019). Correlación: no toda correlación implica causalidad. from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-91902019000300354

Ruiz, L. E. M., & Romero, S. P. L. (2017). Búsqueda de patrones para mejorar productos y servicios en las bibliotecas. INVESTIGACIÓN BIBLIOTECOLÓGICA, Vol. 31, Núm. 72. México. from http://132.248.242.6/~publica/conmutarr.php?arch=3&idx=1056

Ruiz, L. E. M., & Romero, S. P. L. (2018). Resultados Obtenidos En Un Proceso De Minería De Datos Aplicado A Una Base De Datos Que Contiene Información Bibliográfica Referida A Cuatro Segmentos De La Ciencia. *Journal of Information Systems and Technology Management – Jistem USP, Vol. 15, 2018*. doi: 10.4301/S1807-1775201815003

Ruiz Varona, A., Nogueras, I. J., & Lacasta, M. J. (2020). Diseño y aplicación metodológica para la caracterización multidimensional y análisis de la trayectoria del proceso de decrecimiento a nivel municipal en España. from https://hal.archives-ouvertes.fr/hal-03114087/

Salazar, T. J. I., & Girón, C. E. (2021). Análisis y aplicación de algoritmos de minería de datos. from https://revistas.uniminuto.edu/index.php/Pers/article/view/2547/2139

Sánchez, S. P. (2021). Diseño De Una Red Neuronal Convolucional Para La Segmentación De Estructuras Subcorticales Cerebrales from https://riunet.upv.es/handle/10251/169697

Sancho, C. F. (2019a). Redes Neuronales: una visión superficial from http://www.cs.us.es/~fsancho/?e=72

Sancho, C. F. (2020b). Aprendizaje Supervisado y No Supervisado. from http://www.cs.us.es/~fsancho/?e=77

Sancho, C. F. (2021). Aprendizaje Inductivo: Árboles de Decisión. from http://www.cs.us.es/~fsancho/?e=104

Sancho, C. F. (2021a). Mapas Auto-Organizados. from http://www.cs.us.es/~fsancho/?e=76

Sarker, Iqbal H. (2021). Machine Learning: Algorithms, Real‐World Applications and Research Directions. from https://link.springer.com/article/10.1007/s42979-021-00592-x

Schmukler, K. Y. (2017). El enfoque sistémico y sistemático en un proyecto. from https://www.incae.edu/es/blog/2017/01/26/el-enfoque-sistemico-y-sistematico-en-un-proyecto.html

Softoware. (2019). WinIDAMS. Softoware.org. from http://es.softoware.org/apps/download-winidams-for-windows-me-os.html

Sotelo, A. F. I. (2020). PROGRAMA DE ESTUDIO - PROBABILIDADES Y ESTADÍSTICA DESCRIPTIVA E INFERENCIAL - PARA FORMACIÓN DIFERENCIADA. from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwilhfS0qoX0AhWVRTABHVkHAp0QFnoECBkQAQ&url=https%3A%2F%2Fbibliotecadigital.mineduc.cl%2Fbitstream%2Fhandle%2F20.500.12365%2F14318%2Fpr_%2520probs%25203%25C2%25B0%25204%25C2%25B0.pdf%3Fsequence%3D1%26isAllowed%3Dy&usg=AOvVaw1U-GpDFkSsfgvDpGku1WHI

Suárez, M., Milagros de la Caridad , Pérez, U., Aymara , & Jiménez, L., Juan Francisco (2017). Las bibliotecas, sus funciones y retos actuales en Cuba Revista Conrado, 13(58), 7-13. . from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=7&cad=rja&uact=8&ved=2ahUKEwjN5YWf5ePeAhWHuVMKHXhmB1AQFjAGegQIARAC&url=https%3A%2F%2Fconrado.ucf.edu.cu%2Findex.php%2Fconrado%2Farticle%2Fdownload%2F465%2F499%2F&usg=AOvVaw2a0aJSoIkTUIPhd0qMt7oX

SurveyMonkey. (2018). Calculadora del tamaño de muestra. Copyright © 1999 - 2018 SurveyMonkey. from https://es.surveymonkey.com/mp/sample-size-calculator/

Universidad-Barcelona. (2005). MEDIDAS DE ASOCIACIÓN PARA DATOS NOMINALES. from http://www.ub.edu/aplica_infor/spss/cap3-4.htm

Urbizagastegui, R., & Cortes, M. T. (1998). Análisis de citas bibliográficas en la Revista Geológica de Chile. Rev. geol. Chile [online]. from http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0716-02081998000200009&lng=es&nrm=iso

USA, F. R. S. (2018). Annual Performance Plan 2018. Board of Governors of the Federal Reserve System. from https://www.federalreserve.gov/publications/files/2018-gpra-performance-plan.pdf

Utrilla, F. (2018). UNE 166006. Sistemas de Vigilancia e Inteligencia en la Gestión de la I+D+i. La Revista de la Normalización Española. from https://revista.une.org/3/sistemas-de-vigilancia-e-inteligencia-en-la-gestion-de-la-id.html

Valero, M. A. I. (2017). Técnicas estadísticas en Minería de Textos. Universidad de Sevilla. from https://idus.us.es/bitstream/handle/11441/63197/Valero%20Moreno%20Ana%20Isabel%20TFG.pdf?sequence=1

Vargas, G. C. L. (2020). Evaluación del potencial eólico y predicción de la velocidad de viento con Minería de Datos. from https://revista.uisrael.edu.ec/index.php/ro/article/download/368/181

Vázquez, F., Pla, F., & Sánchez, J. S. (2007). Una Propuesta Basada En La Estimación De Las Probabilidades Para La Edición Utilizando El Clasificador K-NN. Universidad Oriente & Universitat Jaume I. from http://marmota.dlsi.uji.es/WebBIB/papers/2007/0_Vazquez-MIA-2007.pdf

Walker, M. G. (1987). How Feasible is Automated Discovery? *IEEE Expert, vol. 2(1), Primavera 1987. Págs 69-82.* from http://www.ksl.stanford.edu/KSL_Abstracts/KSL-86-35.html

Wirth, R., & Hipp, J. (2012). CRISP-DM: Towards a Standard Process Model for Data Mining. Ulm, Germany; Tübingen, Germany. from http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Wirth, R., Shearer, C., Grimmer, U., Reinartz, T., Schlösser, J., Breitner, C., . . . Lindner, G. (2005). Towards process-oriented tool support for knowledge discovery in databases.

European Symposium on Principles of Data Mining and Knowledge Discovery. from https://link.springer.com/chapter/10.1007/3-540-63223-9_123

Yu, S. S., & Chin, F. L. (2021). Applying Educational Data Mining to Explore Viewing Behaviors and Performance With Flipped Classrooms on the Social Media Platform Facebook. University Taiwan. from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.653018/full

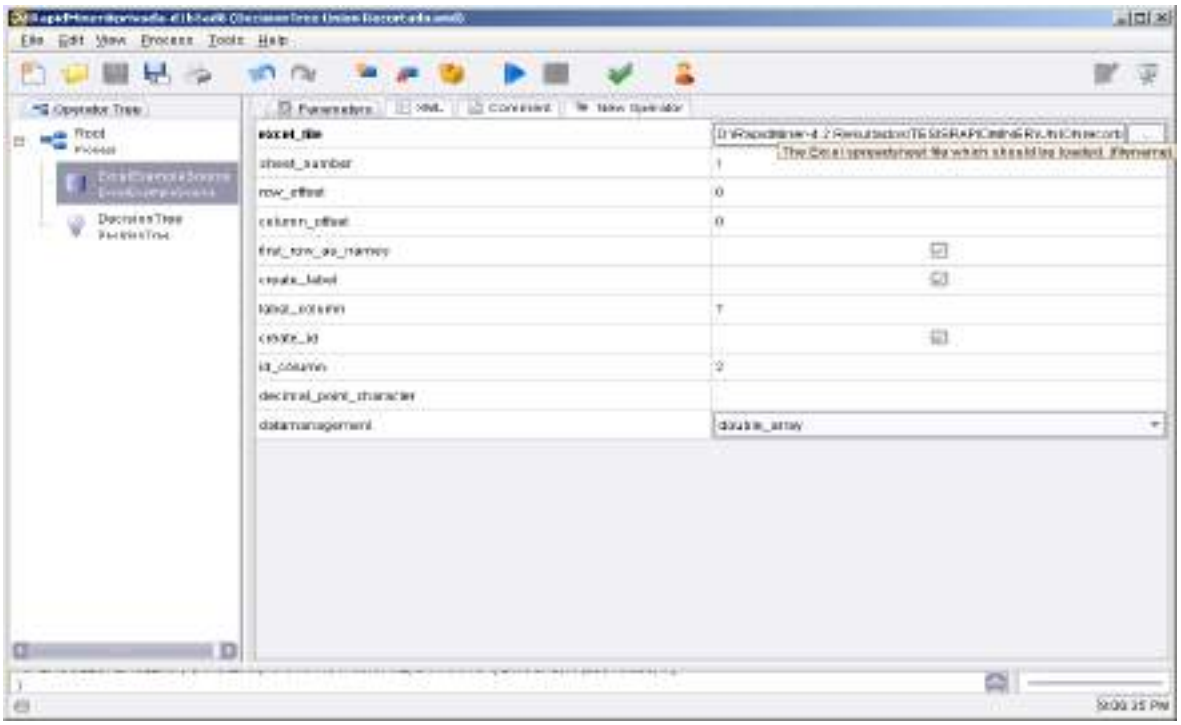
Zambrano, Y., & Cristhian, O. (2021). Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A. . from http://repositorio.ucsg.edu.ec/handle/3317/16543

Zhang, A. (2021). Influence of data mining technology in information analysis of human resource management on macroscopic economic management. from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251483

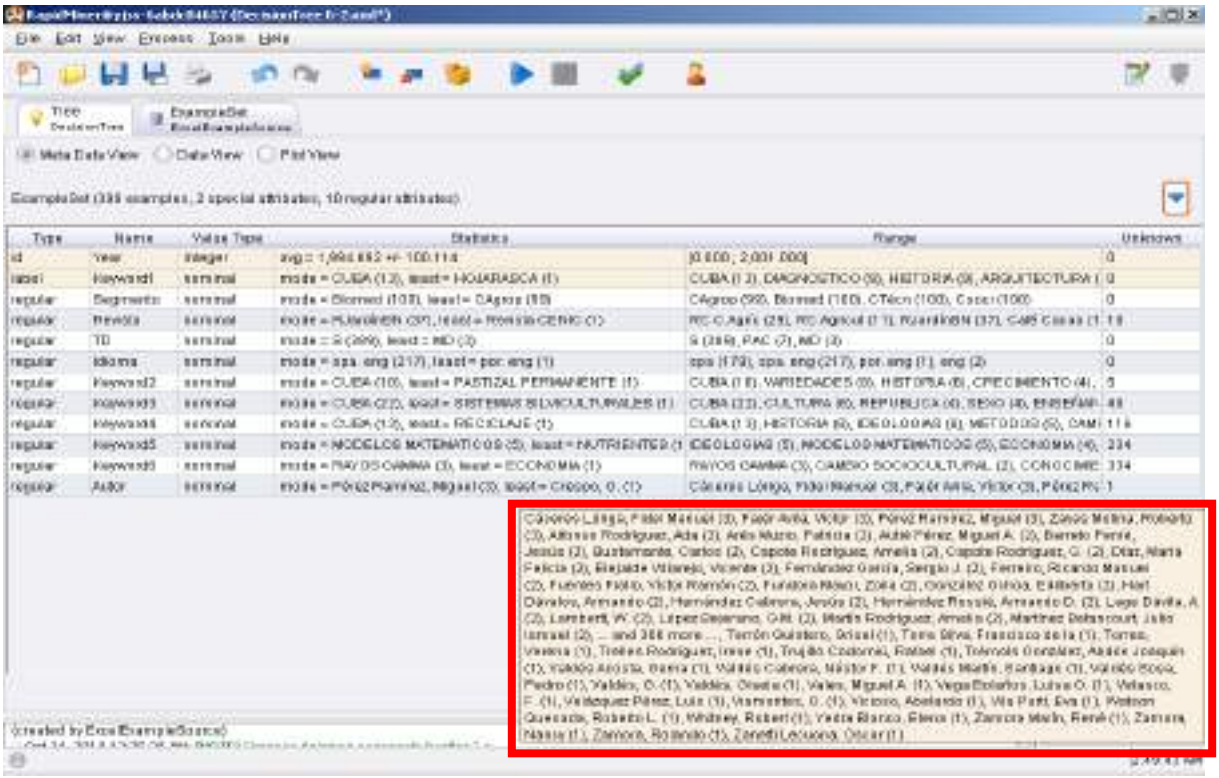
Anexos

Software Rapid Miner V 4.6

Anexo 1. Personalización del software para crear el Árbol de Decisión.

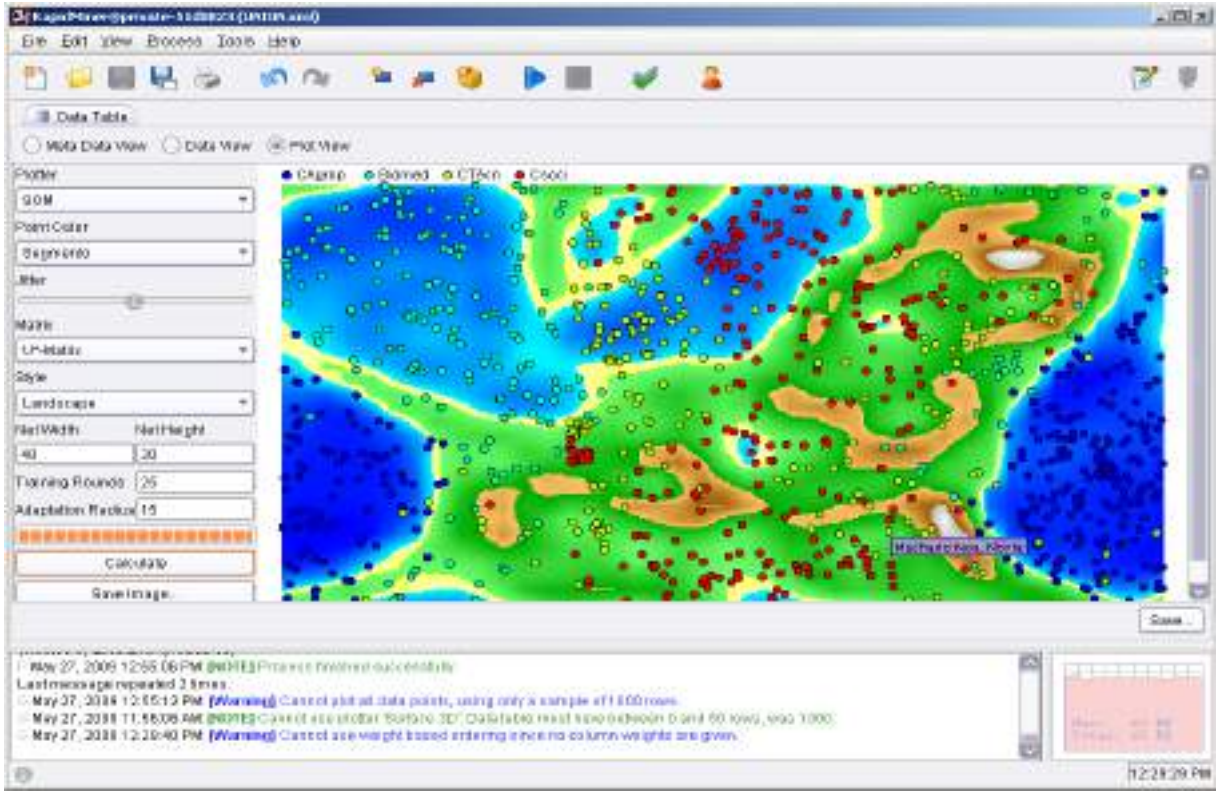


Anexo 2. Resultado de la Meta Data View con ventanas emergentes de Autores

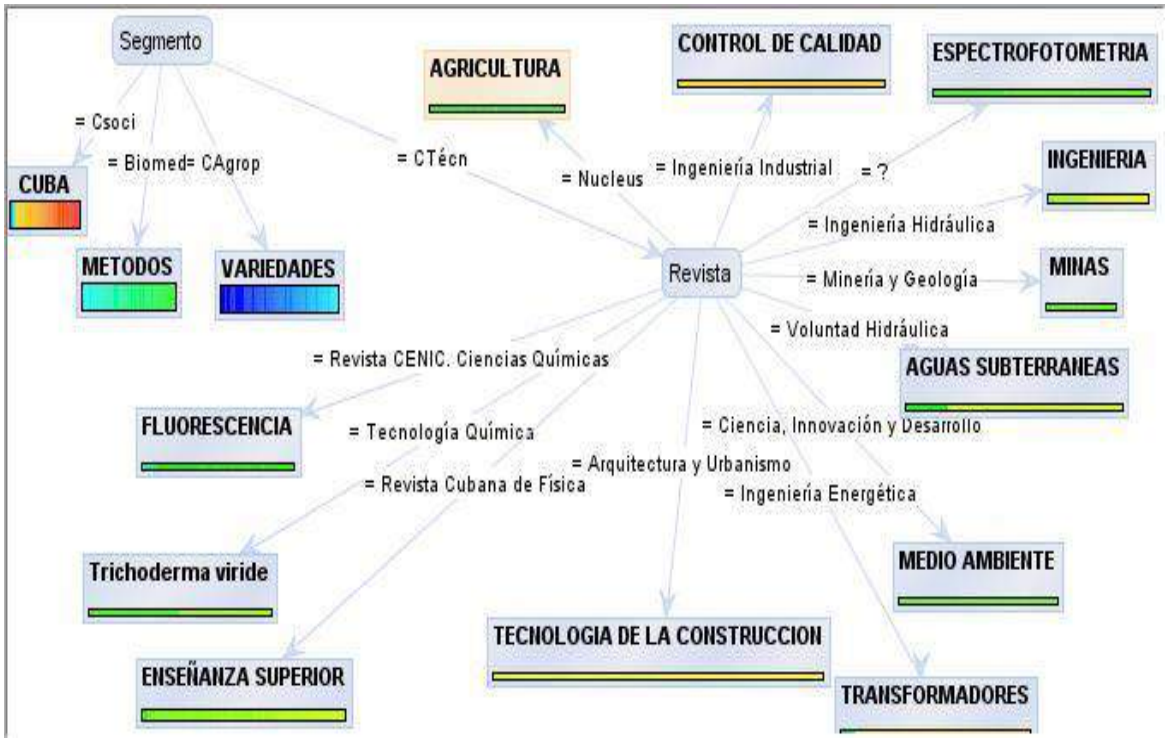


Anexo 3. Resultados de la Meta Data View con Ventana emergente por keyword1.

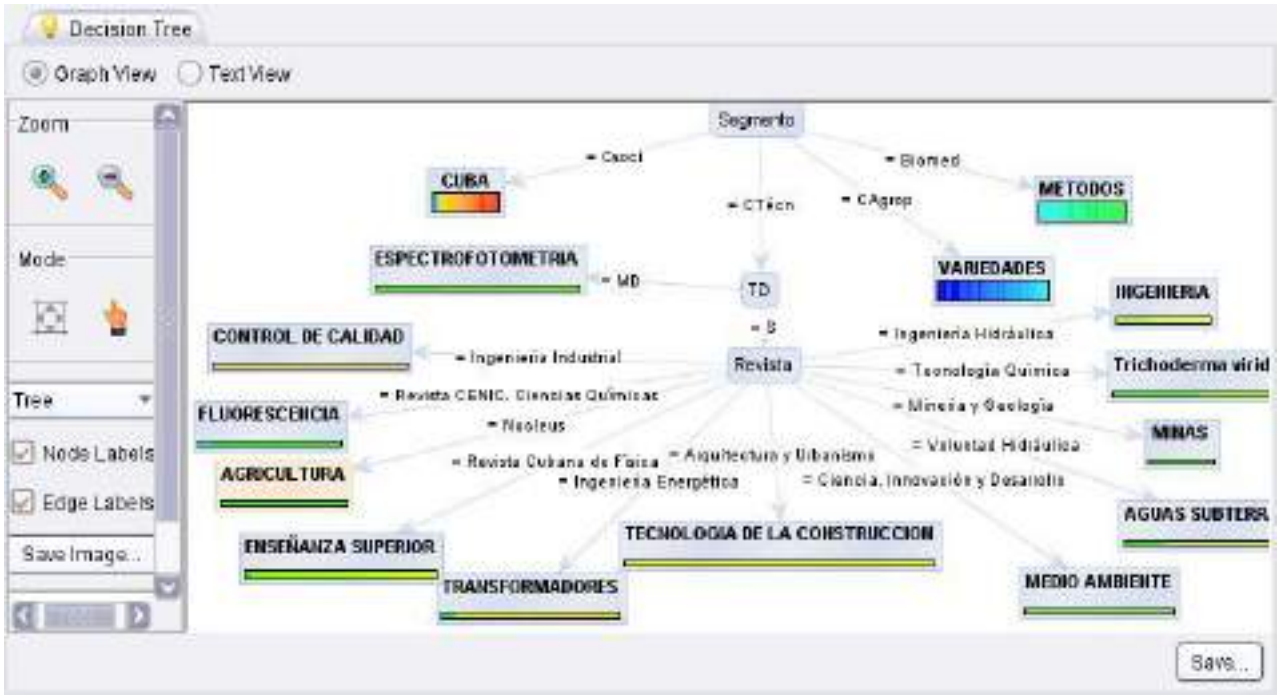
Anexo 4. Resultado del Self Organizing Maps (SOM) x Segmento mostrando Autor 1



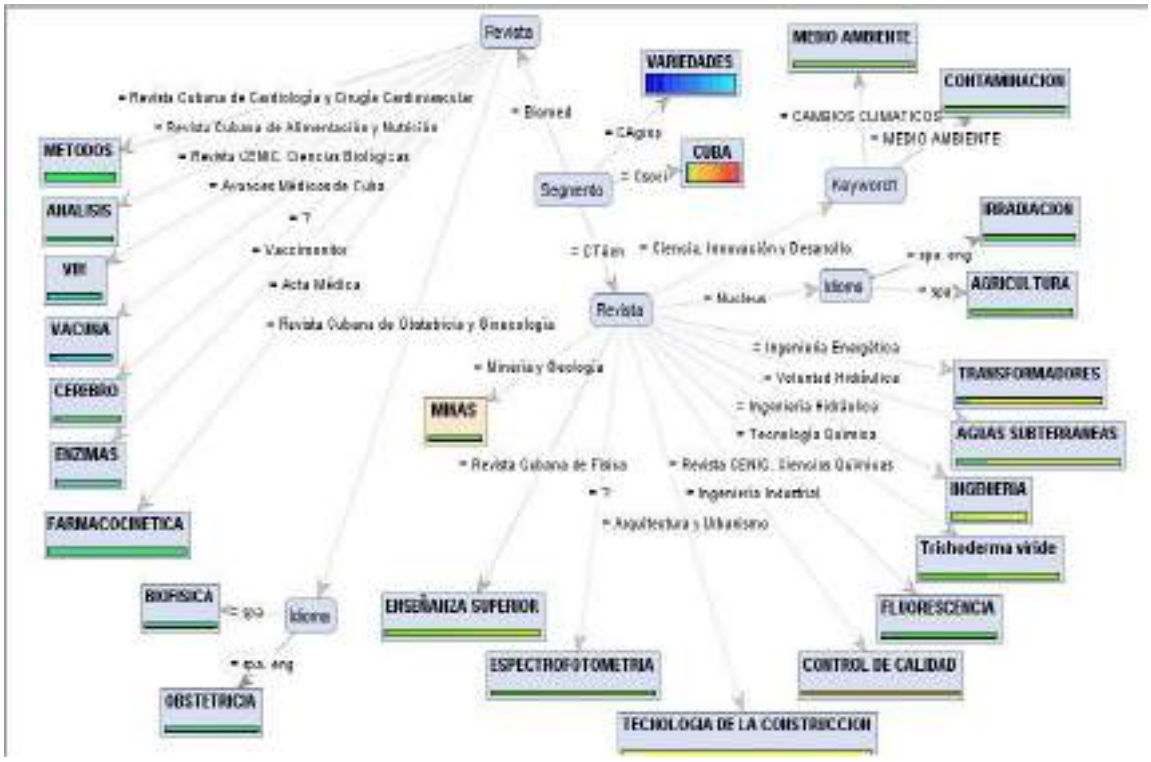
Anexo 5. Resultado de la Primera ejecución del Árbol por Segmento y Keyword1.



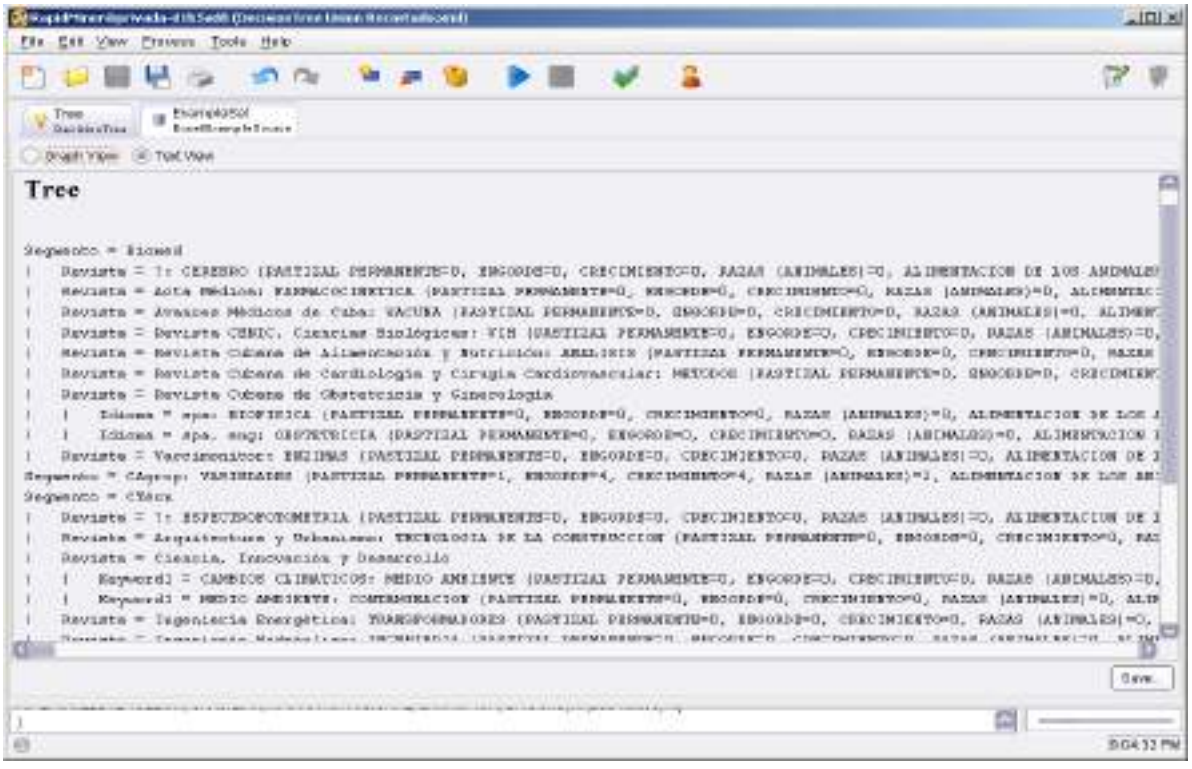
Anexo 6. Resultado de la Segunda ejecución del Árbol por Segmentos y Keyword1 (10491 registros)



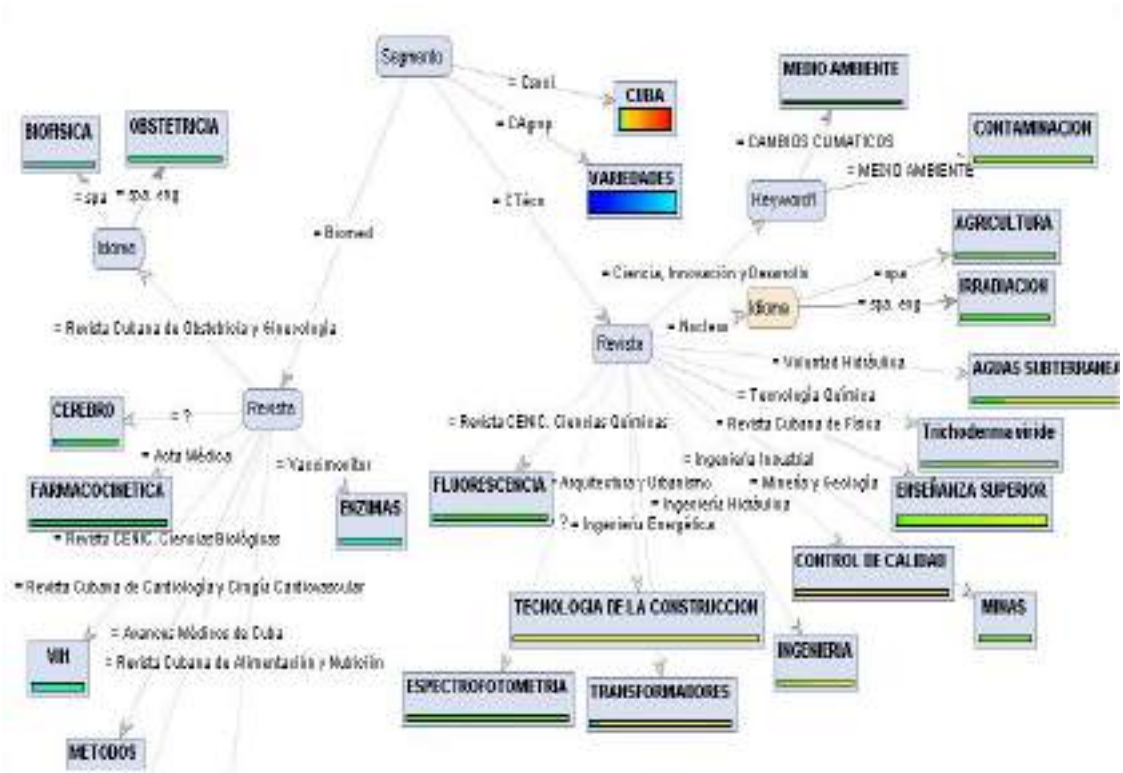
Anexo 7. Resultado de la Tercera ejecución del Árbol por Segmentos y Revistas.



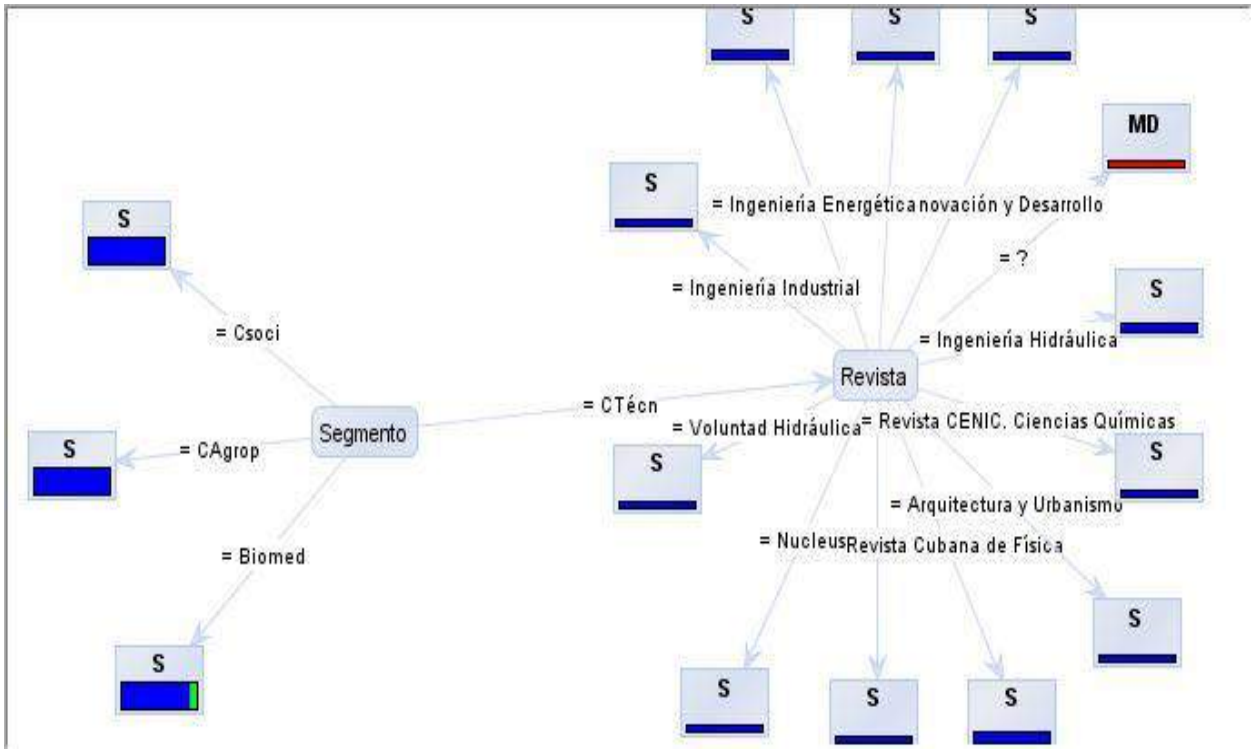
Anexo 8. Resultado del Text View, tercera ejecución del Árbol por Segmentos y Revistas.



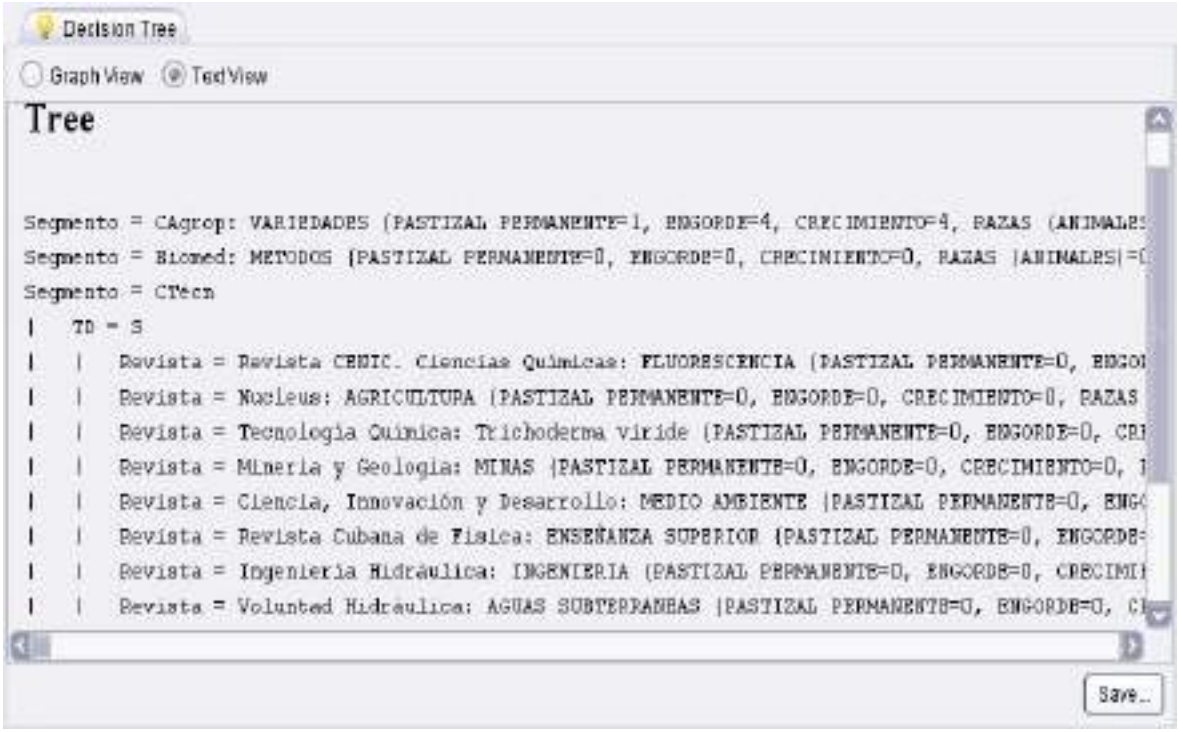
Anexo 9. Resultado del Árbol de Decisión (CHAID7-2) UnionRecortado2



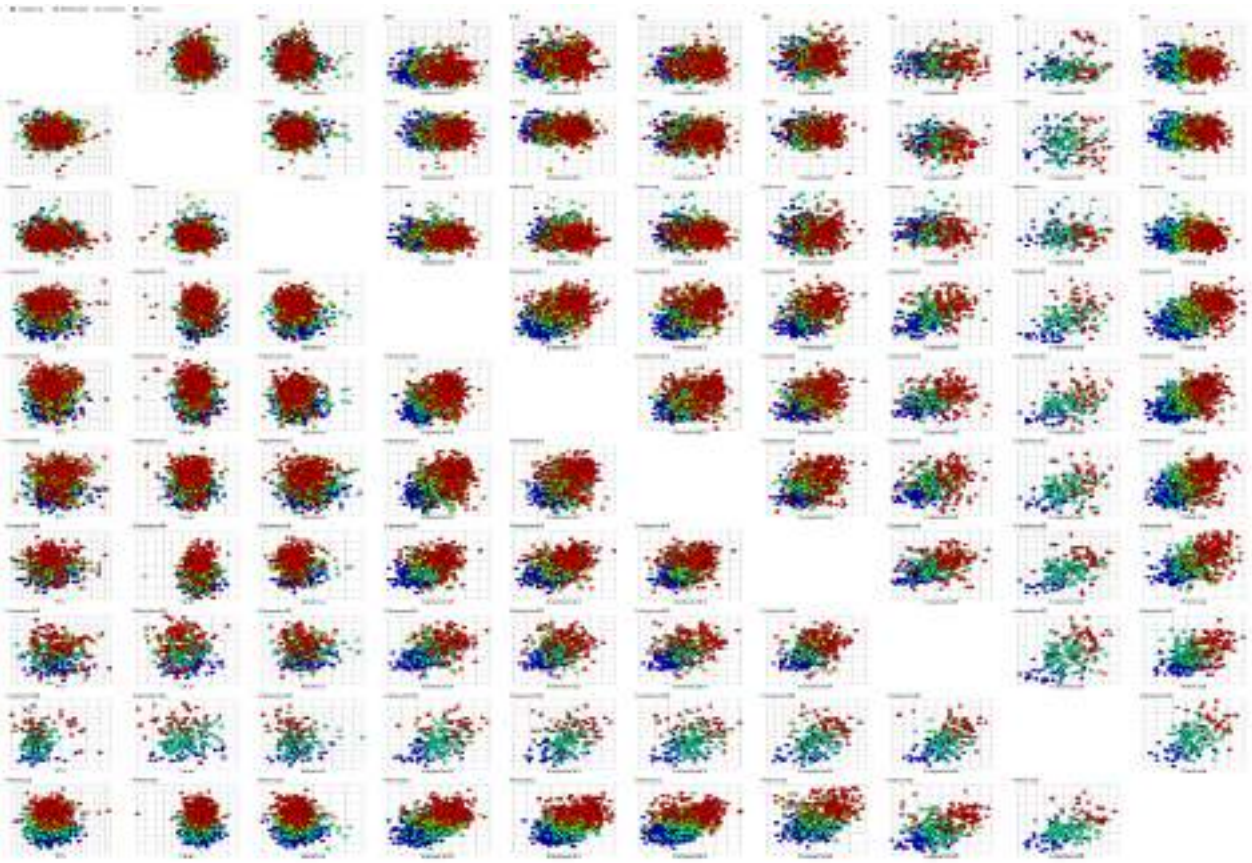
Anexo 10. Resultado del Árbol por Segmentos y tipo de documentos (TD).



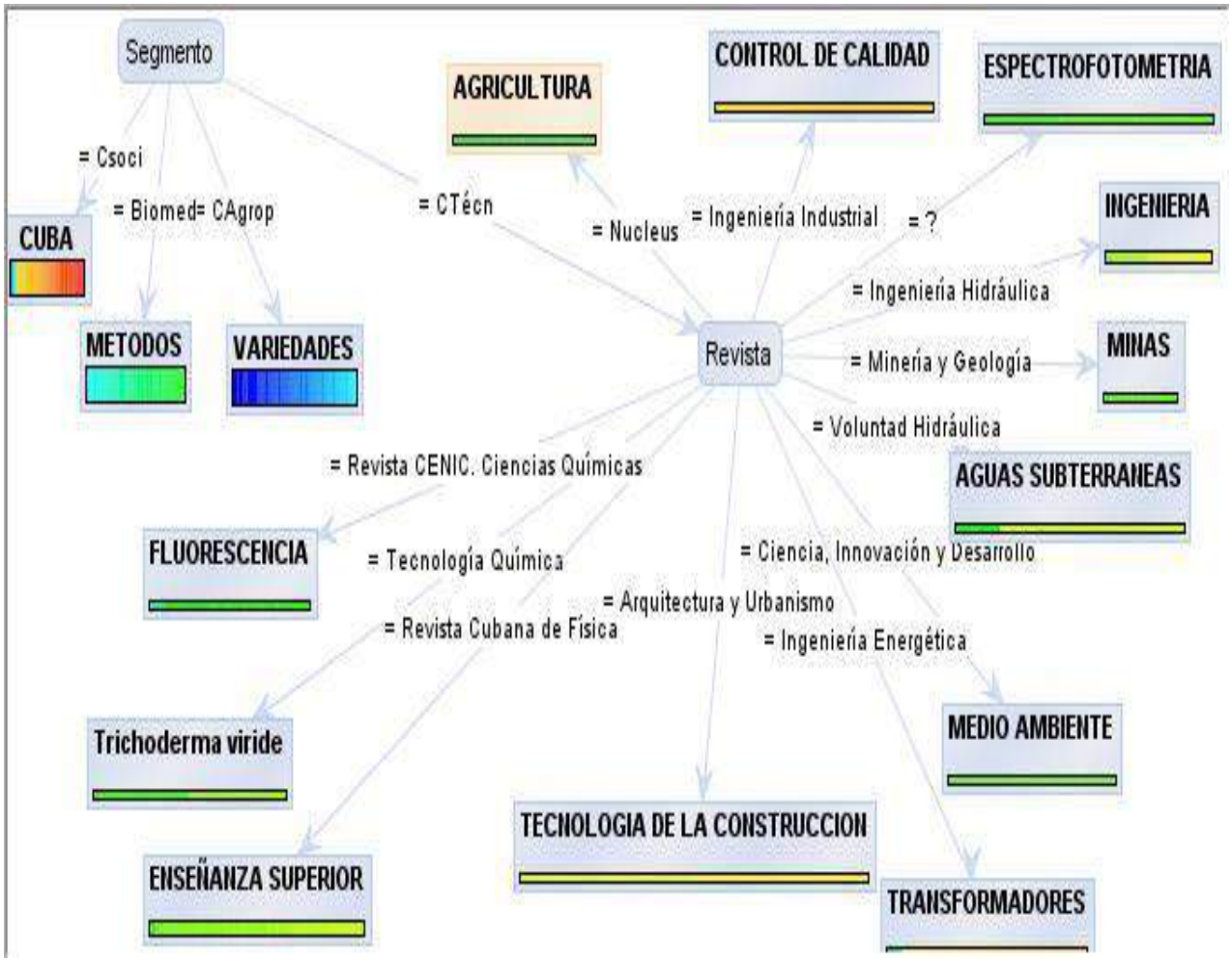
Anexo 11. Resultado del Text View del Árbol por Segmentos y Tipo de documento (TD).



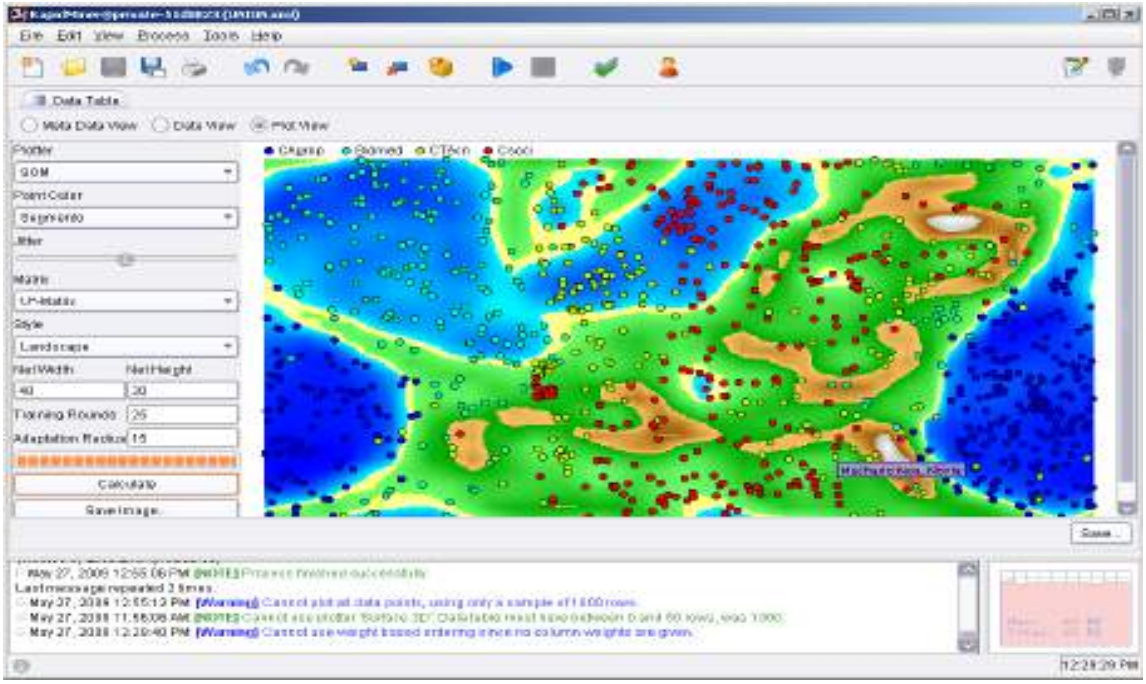
Anexo 12. Resultado del Scatter matriz Plot x Segmento.



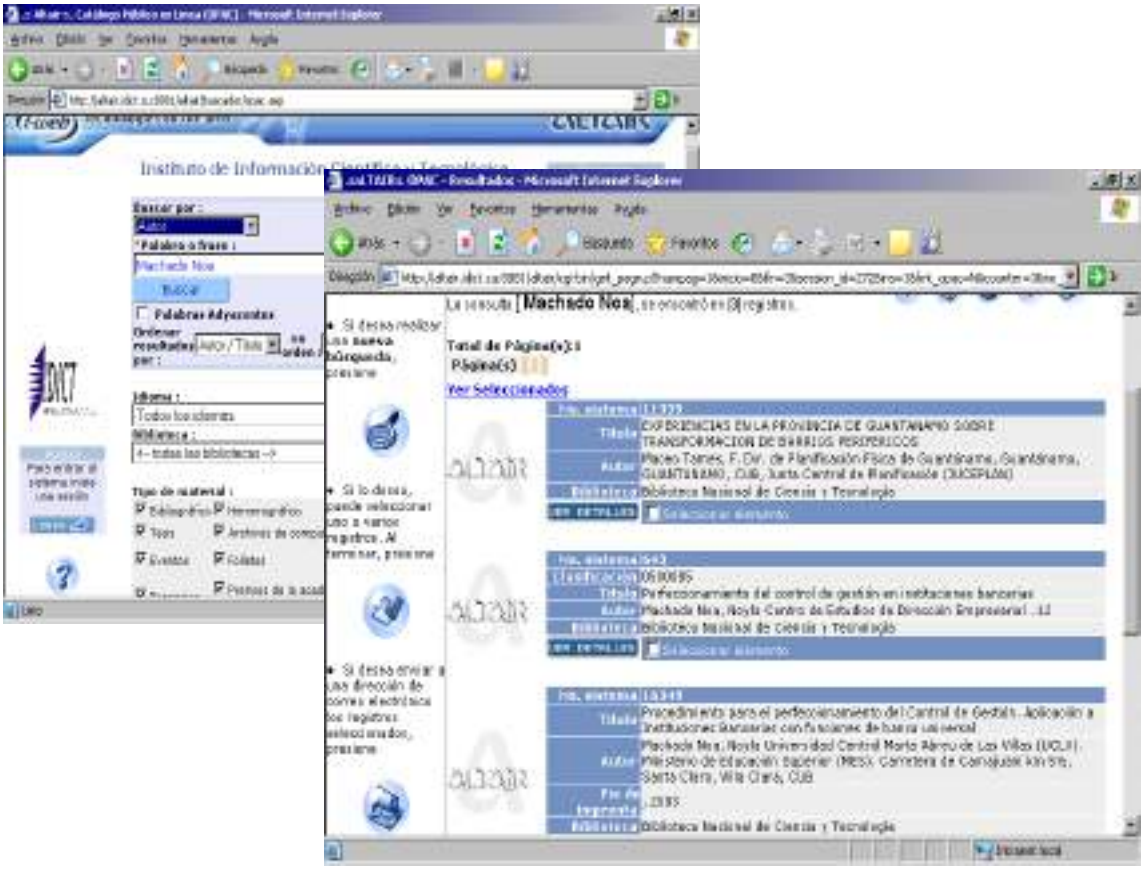
Anexo 13. Resultado del Árbol por Segmento y Keyword1.



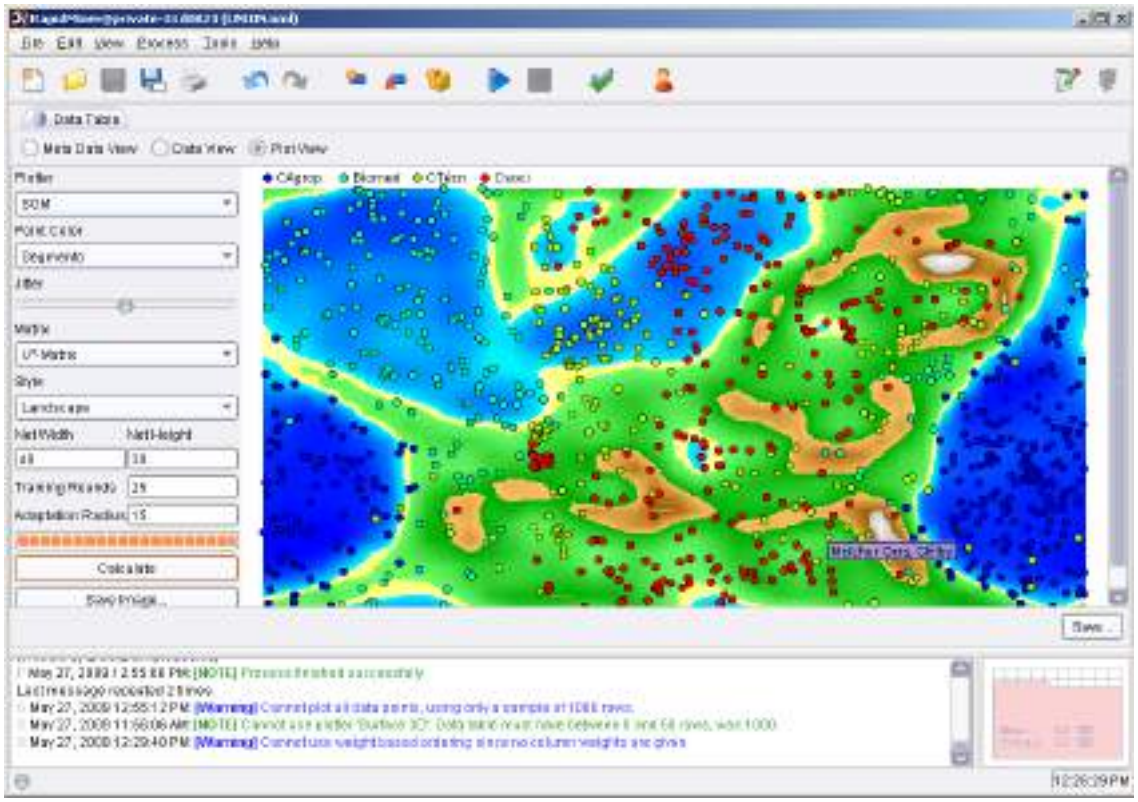
Anexo 14. Resultado del Self Organizing Maps (SOM) x Segmento mostrando Autor 1



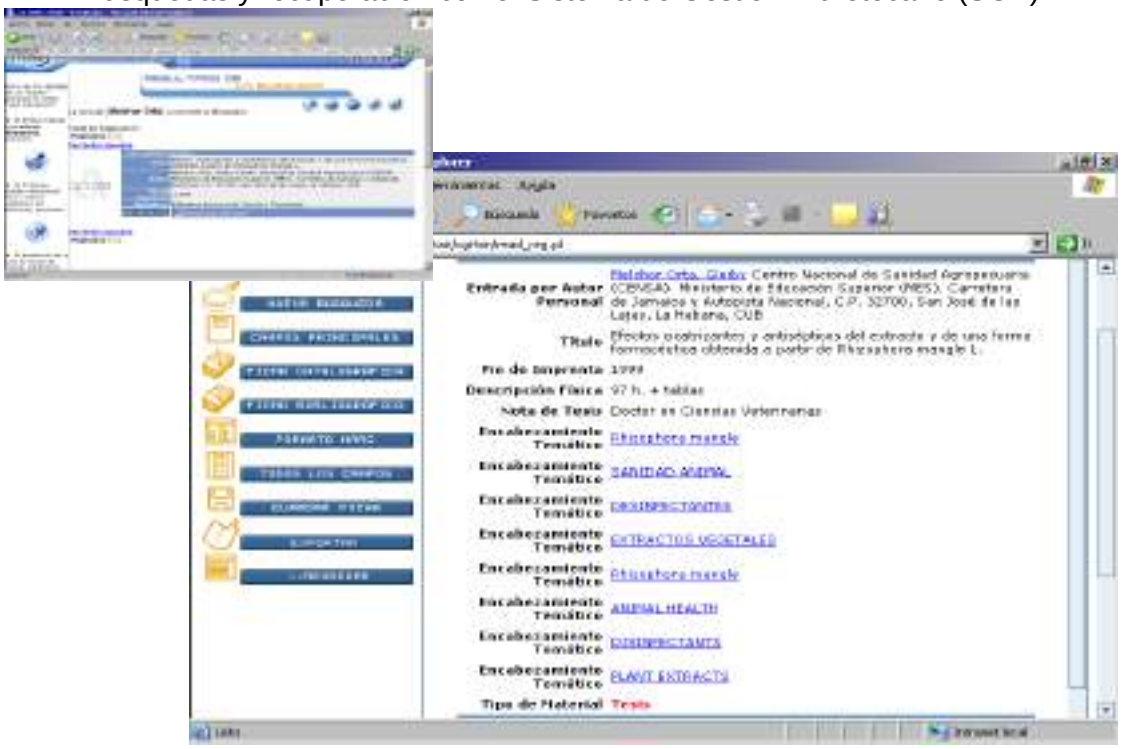
Búsquedas y recuperación con el Sistema de Gestión Bibliotecario (SGB)



Anexo 15. Resultado del Self Organizing Maps (SOM) x Segmento mostrando Autor 2

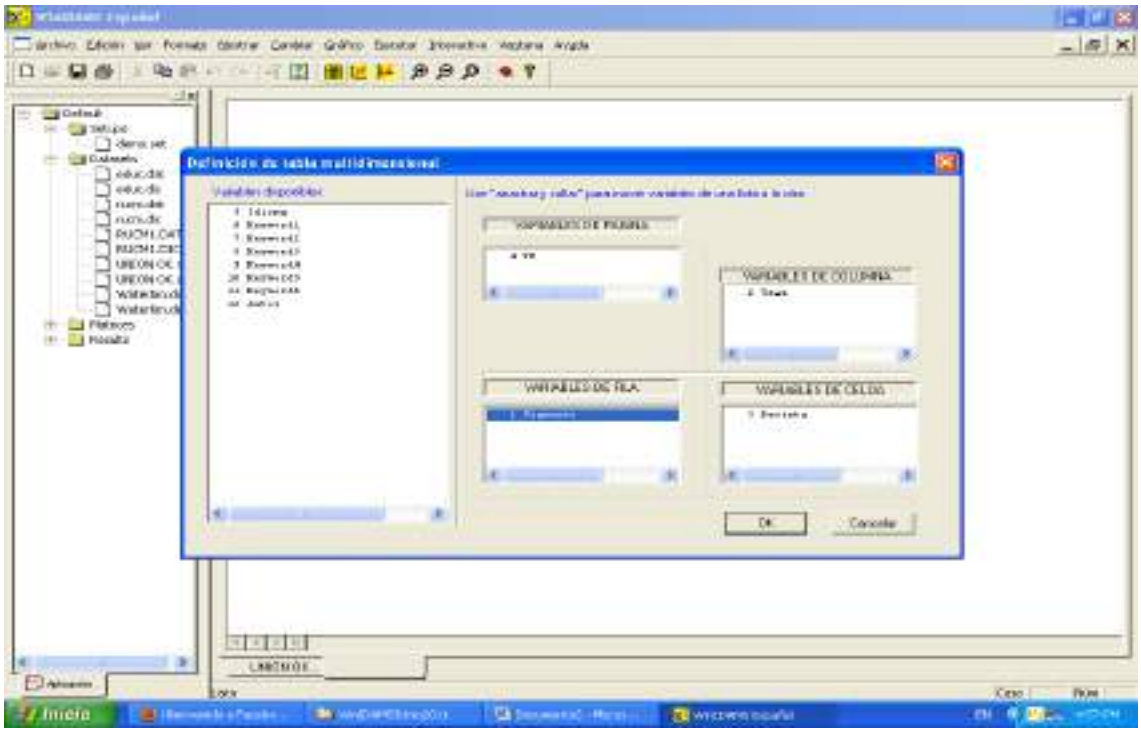


Búsquedas y recuperación con el Sistema de Gestión Bibliotecario (SGB)

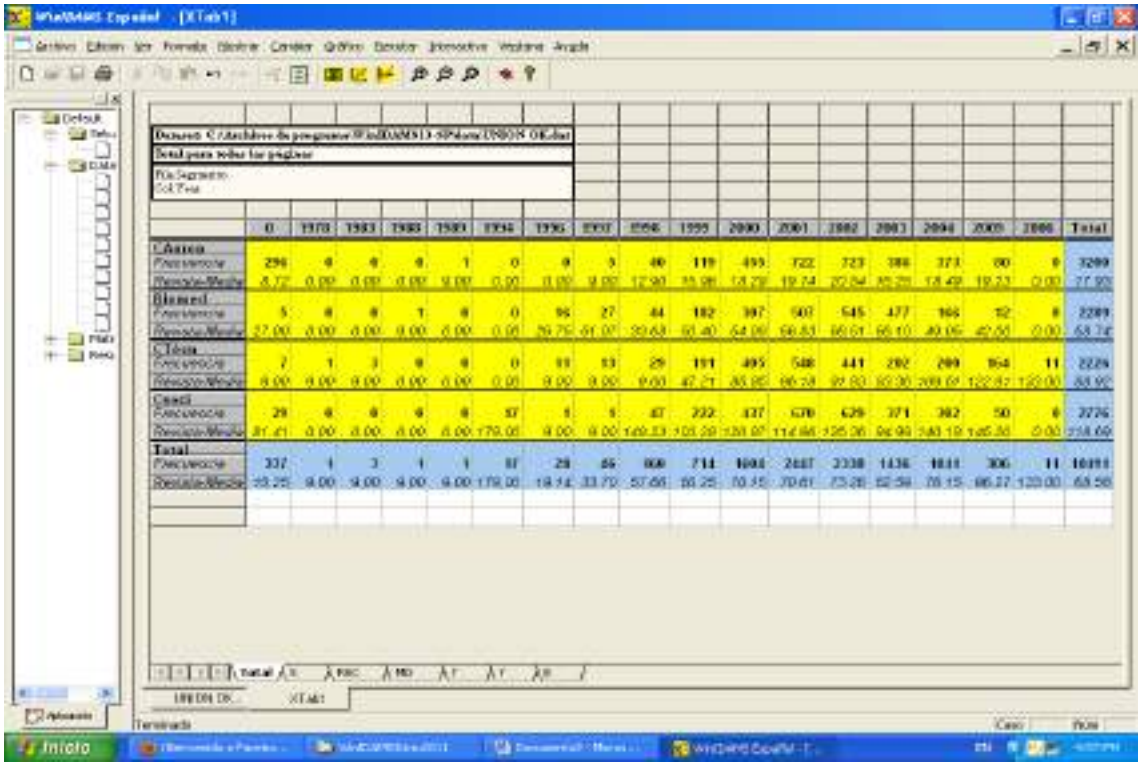


Anexo 16. Gráfico de barras. Distribución por Tipo de Documento (TD) / Segmento.

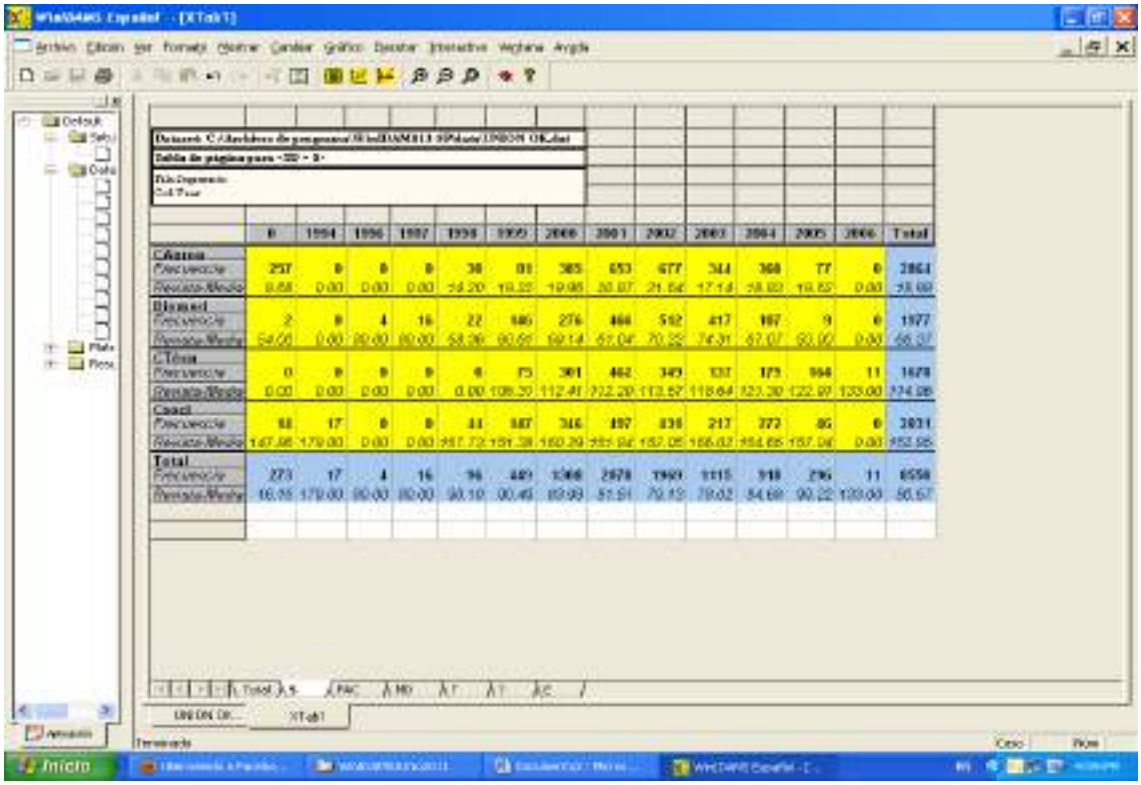
Anexo 17. WinIDAMS. Tablas Multidimensionales



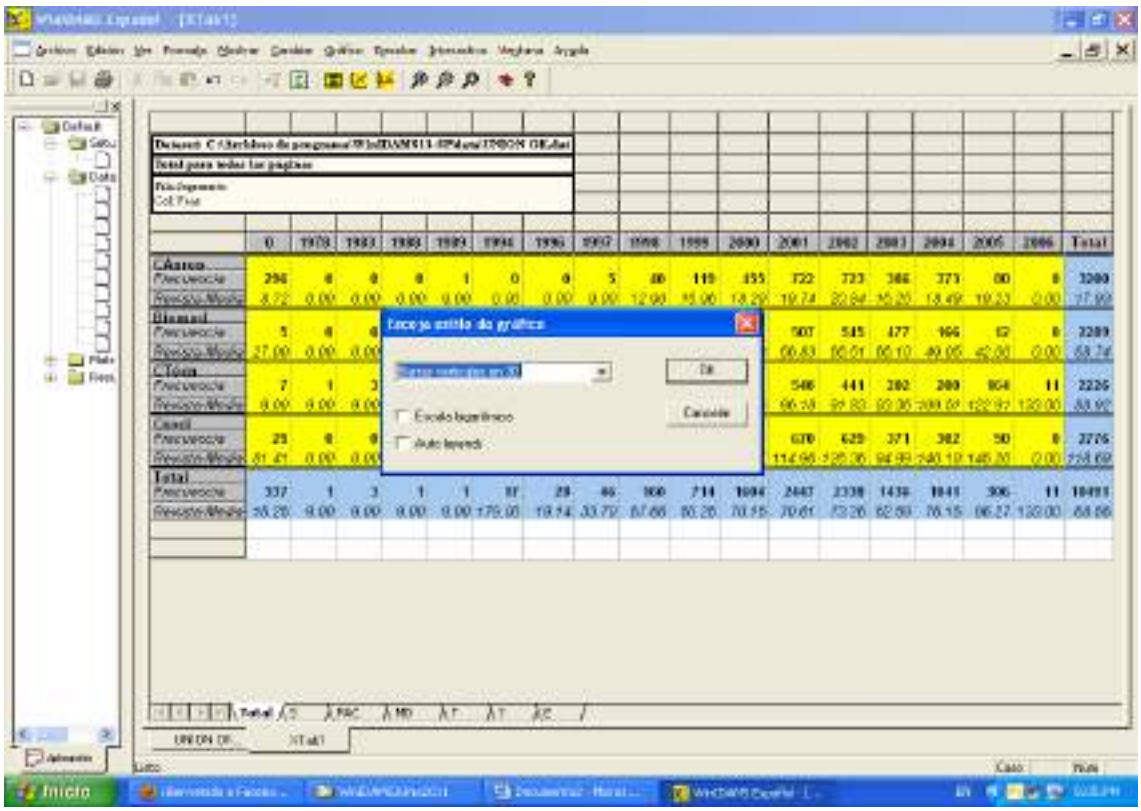
Anexo 18. Resumen Total de las Revistas por Segmento y Año



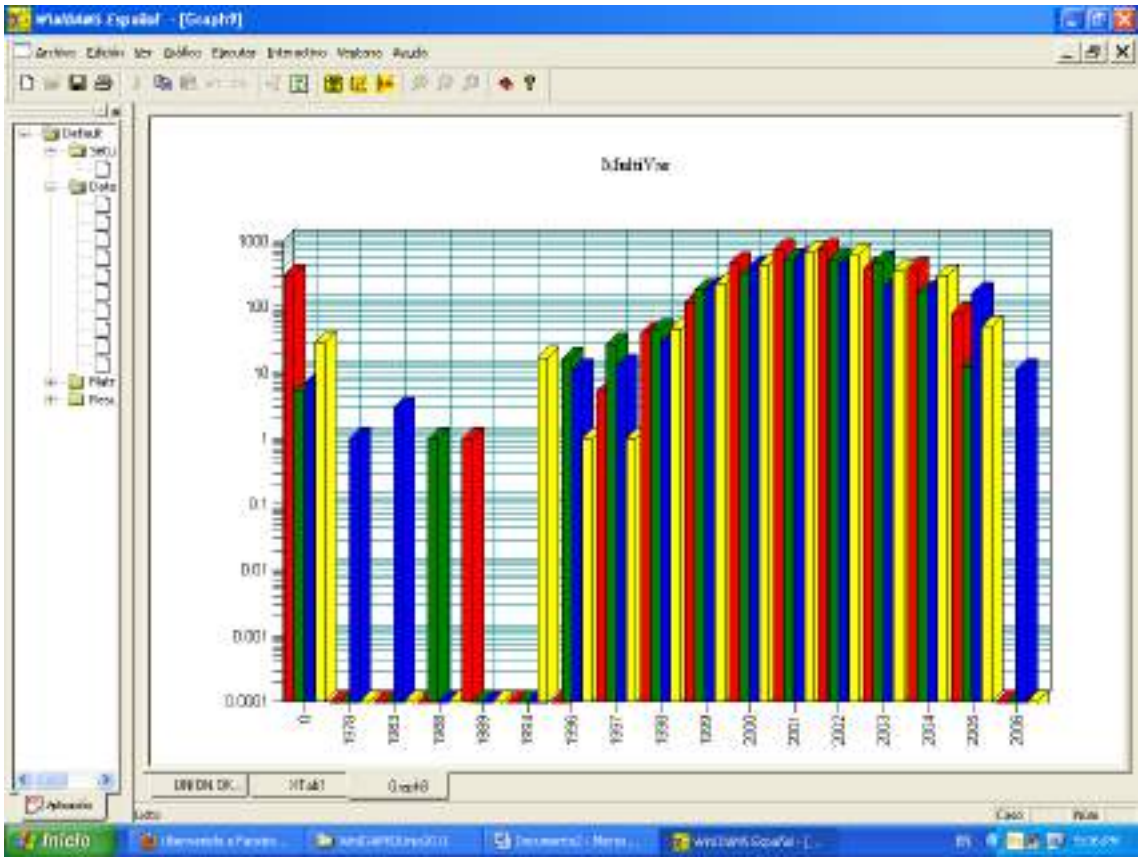
Anexo 19. Resumen en Serie de las Revistas por Segmento y Año.



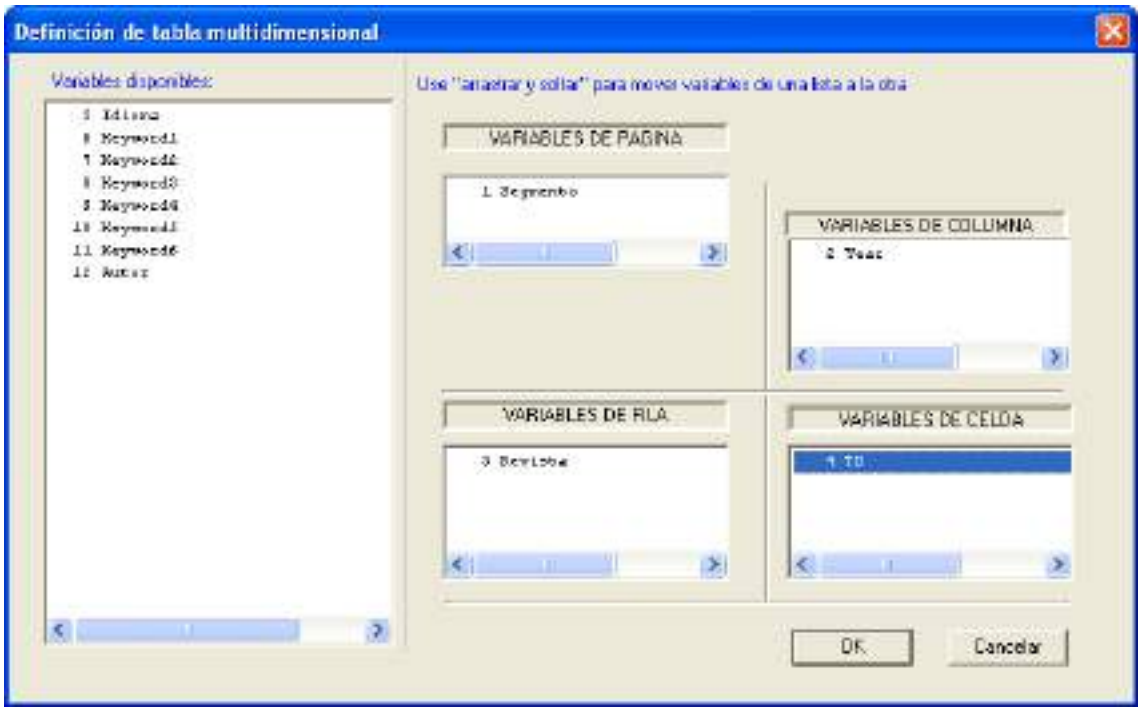
Anexo 20. Preparando gráfico para el resumen total de las Revistas por Segmento y Año



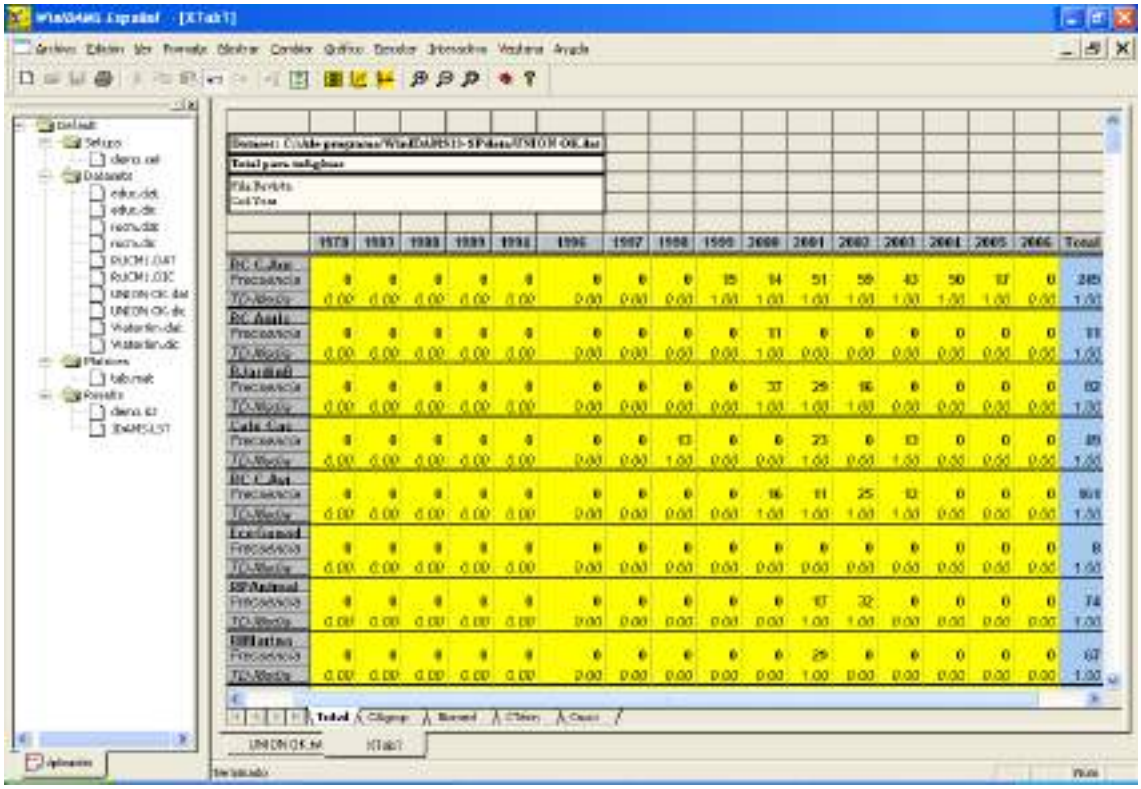
Anexo 21. Gráfico vertical 3D del resumen total de las Revistas por Segmento y Año



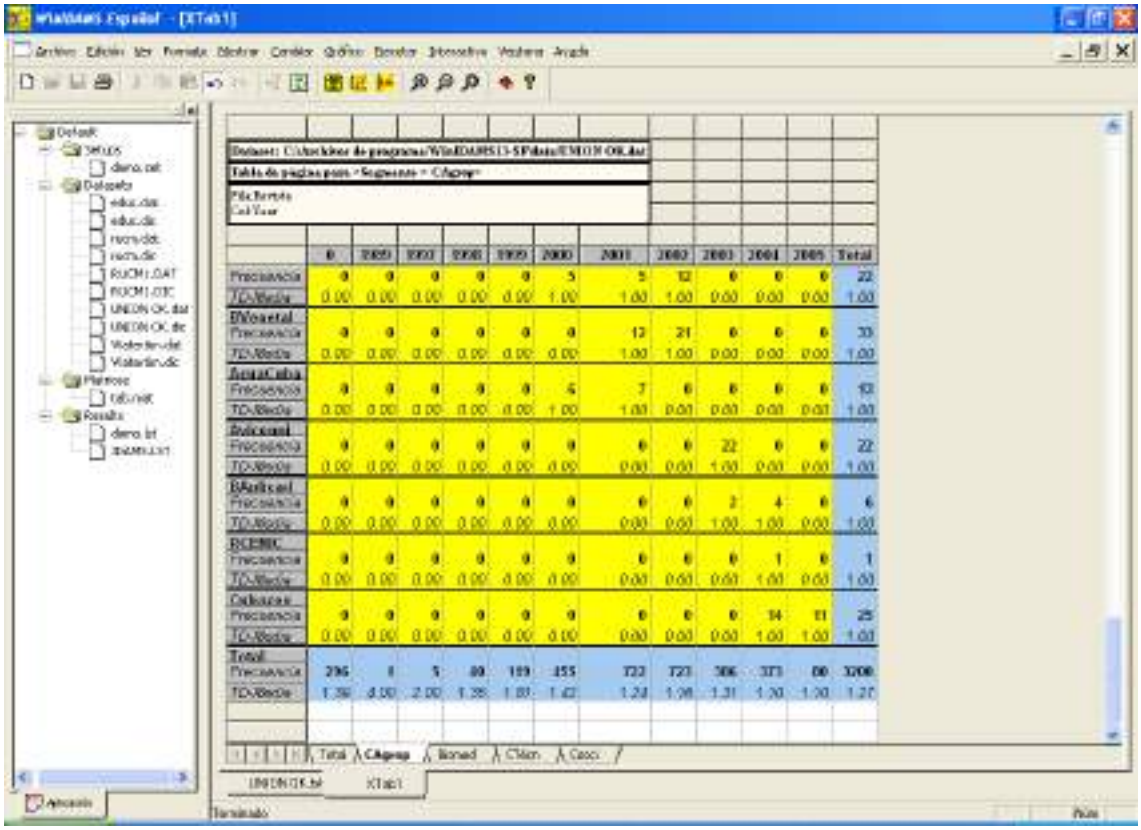
Anexo 22. Definición de nueva tabla multidimensional



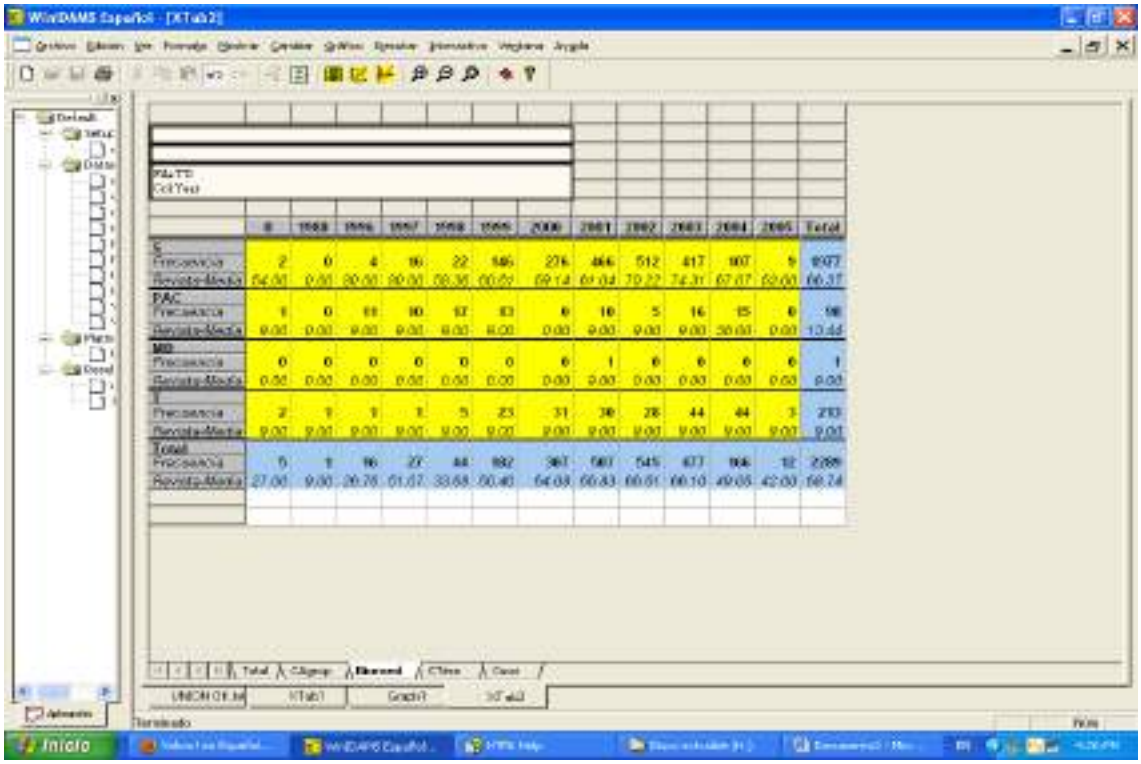
Anexo 23. Resultados del TD por Revista y Año



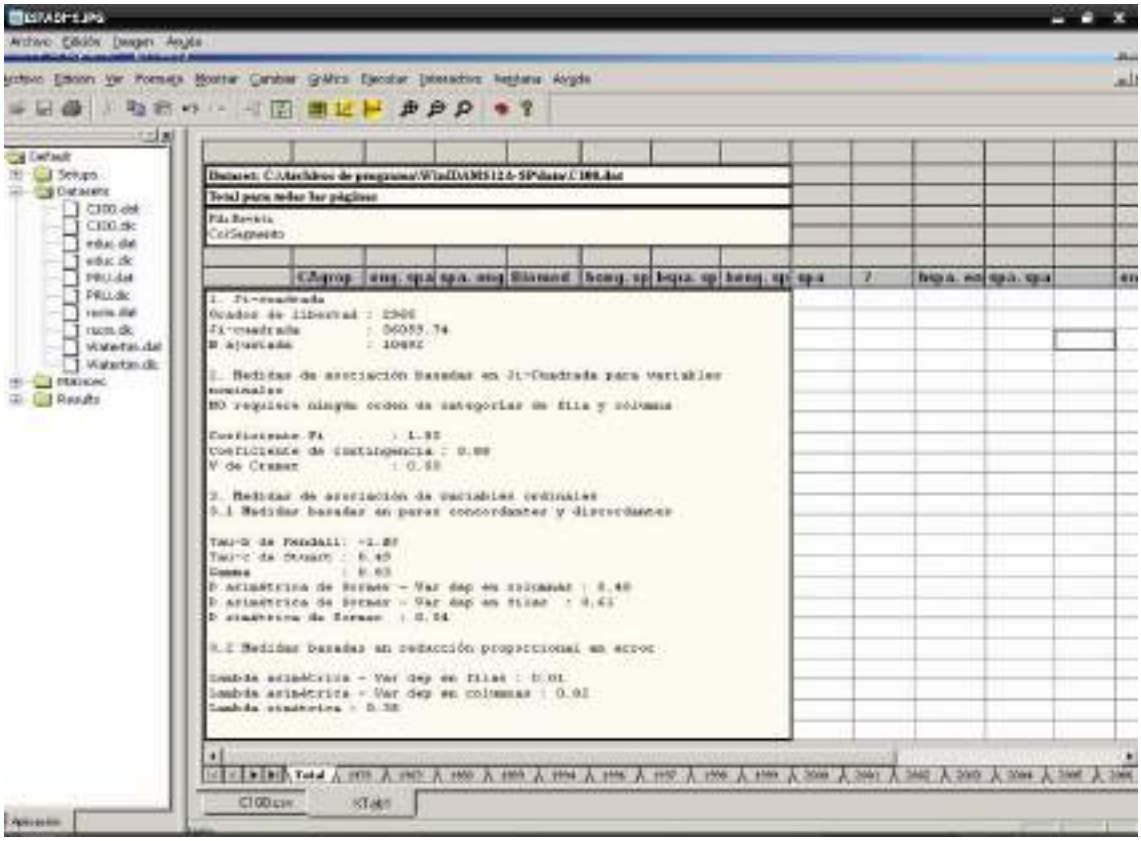
Anexo 24. Final de la Tabla de TD por Revista y Año



Anexo 25. Resumen por Segmento de las Revistas por TD y Año

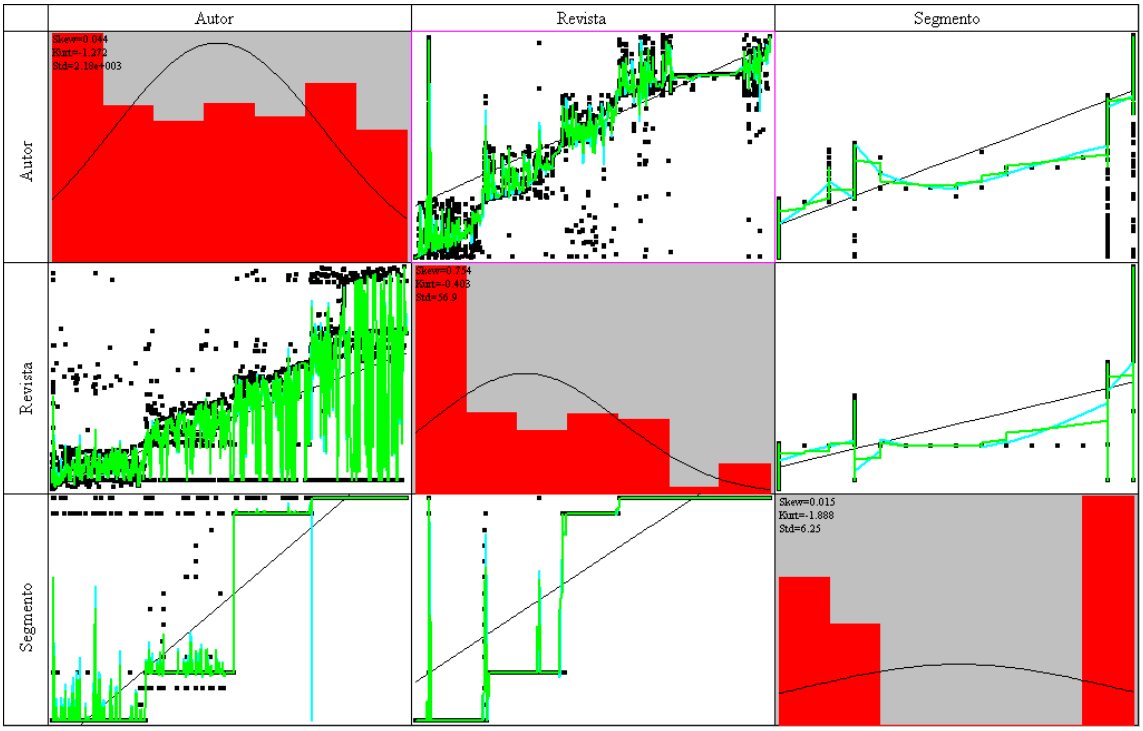


Anexo 26. Estadística con 100 registros y Tabla (Año, Revista, Segmento e Idioma).

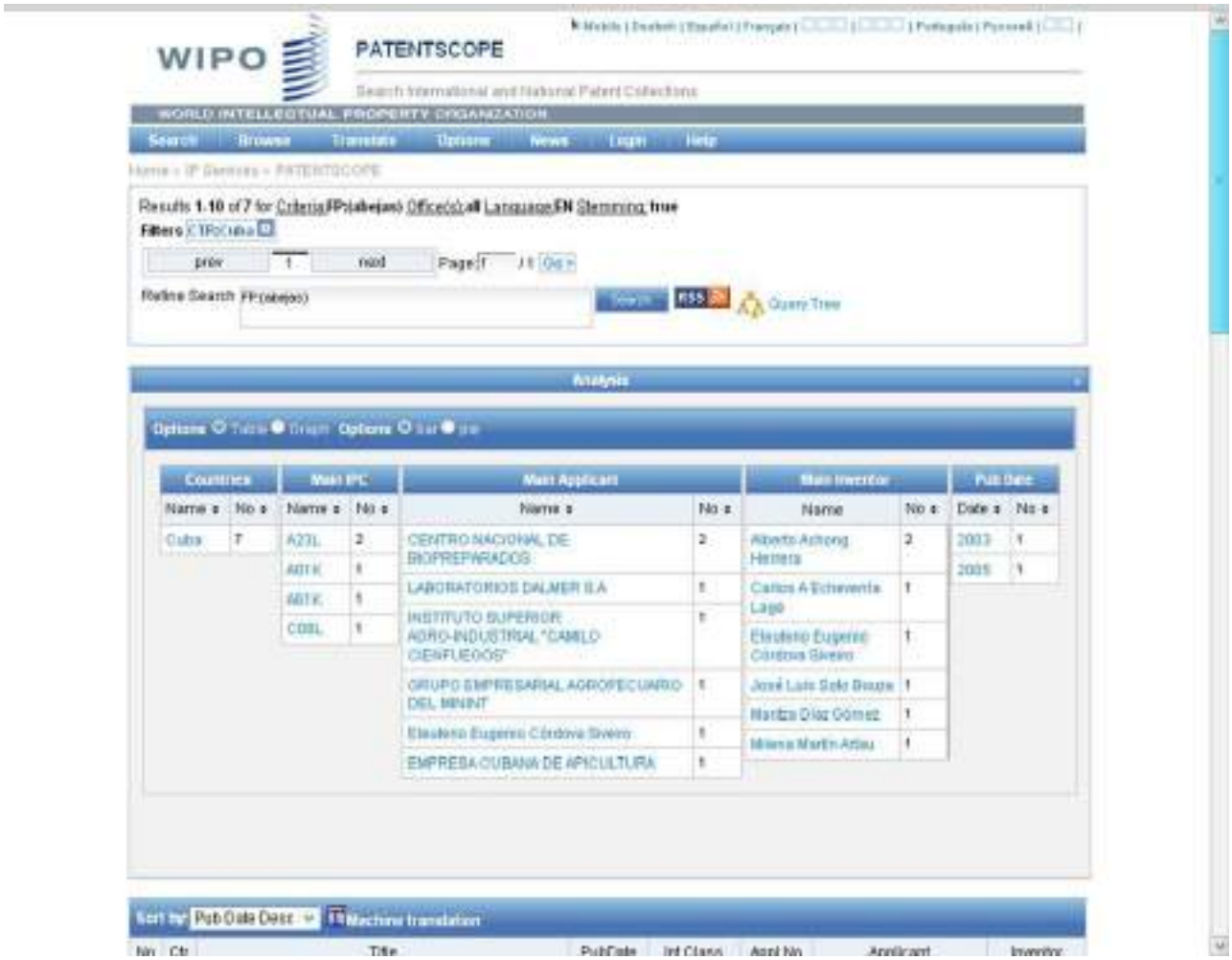


Anexo 27. Exploración gráfica de Datos.

Anexo 28. Histograma con líneas de regresión y correlación



Anexo 29. Búsqueda de patentes de Cuba. Metadato 'Abejas'



Anexo 30. Encuesta realizada al personal bibliotecario para el Control y Monitoreo sobre la satisfacción de la Metodología aplicada.
https://docs.google.com/forms/u/0/

```
FREQUENCIES VARIABLES=V1 V2 V3 V4 V5
  /FORMAT=LIMIT(50)
  /ORDER=ANALYSIS.
```

Frecuencias

```
(Conjunto_de_datos1) G:\PROFE\1 CON ROMERO - TODO NVO\RESULTADOS
ENCUESTA\NVO RESUL SPSS\Datos Fuentes  en SPSS.sav
```

Estadísticos

		V1	V2	V3	V4	V5
N	Válidos	23	23	23	23	23
	Perdidos	0	2	0	1	0

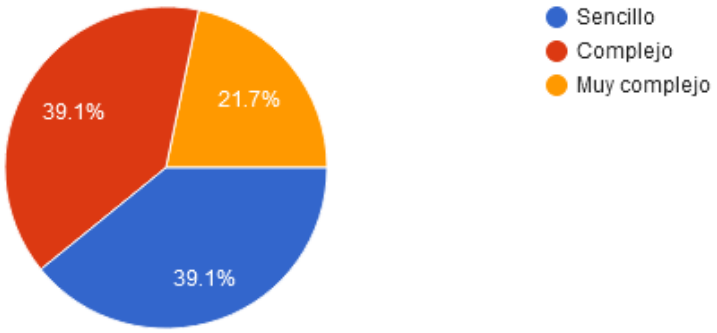
Tablas de frecuencia

V1

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Complejo	9	39.1	39.1	39.1
	Muy complejo	5	21.7	21.7	60.9
	Sencillo	9	39.1	39.1	100.0
	Total	23	100.0	100.0	

El trabajo de selección y preparación de los datos antes de la importación fue?

23 respuestas

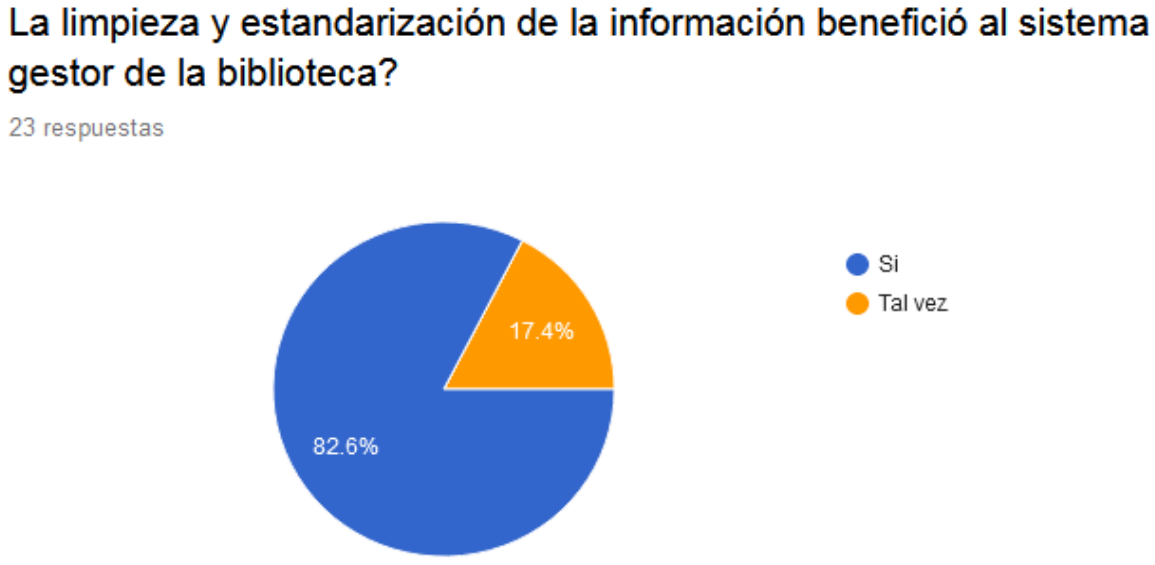


V2

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Si	18	85.7	85.7	85.7
	Tal vez	3	14.3	14.3	100.0
	Total	21	100.0	100.0	

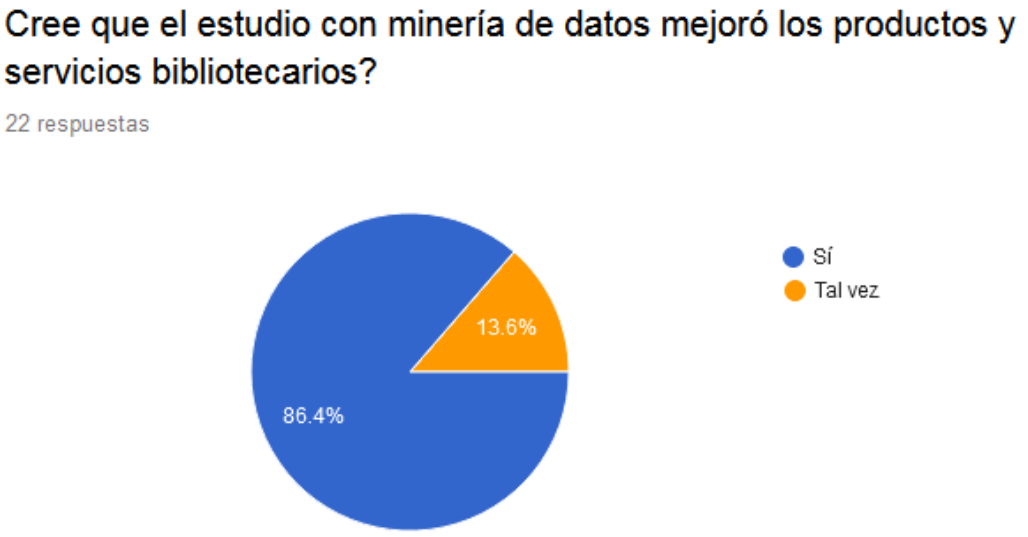
V3

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Si	19	82.6	82.6	82.6
	Tal vez	4	17.4	17.4	100.0
	Total	23	100.0	100.0	



V4

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	19	86.4	86.4	86.4
	Tal vez	3	13.6	13.6	100.0
	Total	23	100.0	100.0	



V5

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Beneficiosa	14	60.9	60.9	60.9
	Buena	8	34.8	34.8	95.7
	Mala	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Por regla de tres: 23 es 100%, como 14 es χ % =

Anexo 31. Encuesta para conocer el nivel de satisfacción del Cliente usando la técnica de Iadov

Cuestionario para conocer el nivel de satisfacción del cliente, por la Metodología creada y aplicada en la gestión de la información, con el uso de la Minería de Datos.

1. Usas con frecuencia los servicios bibliotecario?
   a. Si
   b. No
2. Que es lo que más te gusta de los servicios bibliotecario? _____
3. Que es lo que menos te gusta de los servicios bibliotecarios? _____
4. **Quisieras otros servicios o productos bibliotecarios nuevos?**
   a. No
   b. No sé
   c. Si
5. **Si pudieras aplicar otra Metodología a la información. Lo harías?**
   a. Si
   b. No sé
   c. No
6. **Te gustan los resultados logrados con la nueva Metodología que emplea la Minería de Datos?**
   a. Me gustan mucho
   b. No me gusta tanto
   c. Me da lo mismo
   d. Me disgusta más de lo que me gusta
   e. No me gusta nada
   f. No sé qué decir
7. Como imaginas la biblioteca del futuro? _____
8. Cuanto tiempo dedicas a la lectura? _____

Anexo 32. Respuestas de la encuesta al cliente

1 Si 2 los boletines 3 ir hasta la biblioteca 4 no 5 si 6 a 7 consultar las BDs por internet 8 1 hora por día	1 No 2 los boletines 3 no poder consultar la BD por internet 4 no sé 5 si 6 a 7 alcanzable desde cualquier lugar 8 1 hora por día	1 Si 2 los boletines 3 ir hasta la biblioteca 4 Si 5 si 6 a 7 consultar los catálogos por internet 8 1 hora por día
1 Si 2 los boletines 3 ir hasta la biblioteca 4 no 5 si 6 a 7 consultar los catálogos por internet 8 1 hora por día	1 Si 2 los boletines 3 ir hasta la biblioteca 4 no sé 5 si 6 a 7 consultar las BDs por internet 8 1 hora por día	1 Si 2 los boletines 3 ir hasta la biblioteca 4 Si 5 si 6 a 7 alcanzable desde cualquier lugar 8 1 hora por día
1 Si 2 internet 3 ir hasta la biblioteca 4 no 5 si 6 a 7 alcanzable desde cualquier lugar 8 2 horas por día	1 Si 2 internet 3 ir hasta la biblioteca 4 no sé 5 si 6 a 7 alcanzable desde cualquier lugar 8 2 horas por día	1 Si 2 internet 3 ir hasta la biblioteca 4 Si 5 si 6 a 7 alcanzable desde cualquier lugar 8 2 horas por día
1 Si 2 las búsquedas 3 no poder consultar los catálogos por internet 4 no 5 si 6 a 7 con nuevos productos y servicios 8 5 horas de lectura en internet	1 Si 2 los boletines 3 no poder consultar los catálogos por internet 4 no sé 5 si 6 a 7 con nuevos productos y servicios 8 2 horas de lectura en internet	1 Si 2 las búsquedas 3 no poder consultar los catálogos por internet 4 Si 5 si 6 b 7 con nuevos productos y servicios 8 1 horas de lectura en internet
1 No 2 los boletines 3 ir hasta la biblioteca 4 no 5 si 6 a 7 consultar las BDs por internet 8 2 horas por día	1 Si 2 los boletines 3 ir hasta la biblioteca 4 no sé 5 si 6 a 7 alcanzable desde cualquier lugar 8 3 horas por día	1 No 2 los boletines 3 ir hasta la biblioteca 4 Si 5 si 6 b 7 consultar los catálogos por internet 8 2 horas por día
1 No 2 internet 3 no poder consultar los catálogos por internet 4 no 5 si 6 a 7 alcanzable desde cualquier lugar 8 de 8 a 10 horas por día	1 Si 2 internet 3 no poder consultar los catálogos por internet 4 no sé 5 si 6 a 7 alcanzable desde cualquier lugar 8 de 8 a 10 horas por día	1 Si 2 internet 3 no poder consultar los catálogos por internet 4 Si 5 si 6 b 7 alcanzable desde cualquier lugar 8 de 8 a 10 horas por día
1 No 2 internet 3 ir hasta la biblioteca 4 No 5 no 6 a 7 alcanzable desde cualquier lugar 8 1 hora por día	1 Si 2 los boletines 3 ir hasta la biblioteca 4 no sé 5 no 6 a 7 consultar las BDs por internet 8 1 hora por día	1 Si 2 internet 3 ir hasta la biblioteca 4 Si 5 si 6 b 7 alcanzable desde cualquier lugar 8 1 hora por día
1 Si 2 internet 3 no poder consultar los catálogos por internet 4 No 5 si	1 No 2 internet 3 no poder consultar los catálogos por internet 4 no sé 5 si	1 No 2 internet 3 no poder consultar los catálogos por internet 4 Si 5 si

6 a 7 consultar las BDs por internet 8 de 8 a 10 horas por día	6 a 7 alcanzable desde cualquier lugar 8 de 8 a 10 horas por día	6 b 7 consultar las BDs por internet 8 de 8 a 10 horas por día
1 No 2 internet 3 No 2 internet 3 ir hasta la biblioteca 4 No 5 si 6 a 7 consultar las BDs por internet 8 de 8 a 10 horas por día	1 Si 2 i los boletines 3 No 2 internet 3 ir hasta la biblioteca 4 no sé 5 si 6 a 7 consultar las BDs por internet 8 de 8 a 10 horas por día	1 No 2 internet 3 No 2 internet 3 ir hasta la biblioteca 4 Si 5 si 6 b 7 alcanzable desde cualquier lugar 8 de 8 a 10 horas por día
1 Si 2 internet 3 No 2 internet 3 ir hasta la biblioteca 4 No 5 si 6 a 7 alcanzable desde cualquier lugar 8 de 1 a 2 horas por día	1 Si 2 los boletines 3 No 2 internet 3 ir hasta la biblioteca 4 No sé 5 no 6 a 7 alcanzable desde cualquier lugar 8 de 3 a 5 horas por día	1 Si 2 internet 3 No 2 internet 3 ir hasta la biblioteca 4 Si 5 si 6 a 7 alcanzable desde cualquier lugar 8 de 1 a 2 horas por día

Anexo 33.- Cuadro lógico de Iadov

	4. Quisieras otros servicios o productos bibliotecarios nuevos?								
	No			No sé			Si		
6. Te gustan los resultados logrados con la nueva Metodología que emplea la Minería de Datos?	**5. Si pudieras aplicar otra Metodología a la información. Lo harías?**								
	Si	No sé	No	Si	No sé	No	Si	No sé	No
Máxima satisfacción	9		1	8		2	4		
Más satisfecho que insatisfecho							6		
No definida									
Más insatisfecho que satisfecho									
Máxima insatisfacción									
Contradictoria									

Índice de satisfacción grupal $ISG = \frac{A(+1) + B(+0,5) + C(0) + D(-0,5) + E(-1)}{N}$

$$(ISG) = \frac{24(+1) + 6(+0.5) + 0(0) + 0(-0.5) + 0(0)}{30}$$

$$= \frac{24 + 3 - 0}{30} = \frac{27}{30} = 0.90$$

+1	Máximo de satisfacción
0,5	Más satisfecho que insatisfecho
0	No definido y contradictorio
- 0,5	Más insatisfecho que satisfecho
-1	Máxima insatisfacción

El **Índice de satisfacción grupal es 0.90**, por lo tanto se considera **Máxima satisfacción**