ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Maestría en Sistemas de Información Gerencial

“IMPLEMENTACIÓN DE UNA APLICACIÓN DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE PERFILES DE USUARIOS Y PATRONES DE CONSULTA EN UN CENTRO DE RECURSOS BIBLIOGRÁFICOS”

TRABAJO DE TITULACIÓN

Previo a la obtención del título de:

MAGISTER EN SISTEMAS DE INFORMACIÓN GERENCIAL

Autor:

Jefferson Stalyn Alejandro Domínguez

Guayaquil – Ecuador

2018

AGRADECIMIENTO

Agradezco a Dios por la sabiduría paciencia y sobre todo por la fuerza que ha generado en mí para poder llevar a cabo la culminación de este grado académico, a mi esposa, mi hija, a mi familia en general por la paciencia y el amor recibido; al Ing. Fausto Correa por su colaboración y guía recibida.

DEDICATORIA

Este trabajo se lo dedico a Dios, a la institución a la cual represento y en la que colaboro para su mejoramiento, a las diferentes autoridades por su apoyo; a mis compañeras de trabajo por su colaboración y en general a todos quienes han hecho posible este sueño.

TRIBUNAL DE SUSTENTACIÓN

_____

Mgs. Lenin Freire Cobo

DIRECTOR MSIG

_____

Mgs. Fausto Correa Almazán

DIRECTOR DEL PROYECTO DE GRADUACIÓN

_____

Mgs. Omar Maldonado Dañin

MIEMBRO DEL TRIBUNAL

DECLARACIÓN EXPRESA

La responsabilidad del contenido de este Trabajo de Titulación me corresponde exclusivamente; el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral.

(Reglamento de exámenes y títulos profesionales de la ESPOL)

_____

Jefferson Stalyn Alejandro Domínguez

RESUMEN

Las instituciones de Educación Superior están obligadas a realizar de forma periódica evaluaciones sobre el uso de sus colecciones, que permitan identificar falencias o necesidades de información, y que, a su vez, cumplan con los indicadores de acreditación institucional.

Mediante la utilización de procesos de minería de datos, se ha construido un modelo de reglas de asociación. Reglas que ayudarán en la identificación de perfiles de usuarios, así como, en la identificación de patrones de consulta.

Se ha recopilado datos, producto de las transacciones realizadas para los préstamos de material bibliográfico, correspondientes a los 2 últimos años, las reglas de asociación están basadas en el algoritmo *FP-Gwoth*, considerando las diferentes medidas de interés como *soporte*, *confianza* y *lift*.

Para la implementación, se eligió como herramienta el software Rapidminer, en mismo que dispone de distintos operadores para desarrollar las diferentes etapas de un modelo de minería de datos, iniciando con la carga de datos, realización del proceso ETL, la creación del algoritmo y posterior creación de las reglas de asociación.

Finalmente, se ha analizado algunas relaciones en los resultados de salida, además de emitir una recomendación que permita que la institución reciba de forma periódica esta información, con la finalidad de tomar mejores decisiones para la prestación de servicios hacia los usuarios y para generar acciones que permitan un desarrollo equilibrado de sus colecciones.

***Palabras Clave**: Minería de datos, Algoritmo FP-Growth, Rapidminer, proceso ETL, reglas de asociación*

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURA Y SIMBOLOGÍA

CART	Classification and Regression Trees
CEAACES	Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior
CIB	Centros de Información Bibliográficos
CRAI	Centros de Recursos para el Aprendizaje y la Investigación
CRUD	Create, Read, Update, Delete
DM	Data mining
DW	Data Warehouse
ETL	Extract, Transform and Load
IA	Inteligencia Artificial
ICC/ESOMAR	International Chamber of Commerce / European Society for Opinion and Marketing Research
ID	Identification

ID3	Induction Decision Trees
IES	Instituciones de Educación Superior
IFLA	International Federation of Library Associations and Institutions
JSON	JavaScript Object Notation «notación de objeto de JavaScript»
KDD	Knowledge Discovery in Databases
LOES	Ley Orgánica de Educación Superior
OLAP	On-Line Analytical Processing
SPSS	Statical Package for the Social Sciences
UNESCO	United Nations Educations, Scientific and Culture Organization
VCM	View Conceptual Minable
VOM	View Operative Minable
WWW	World Wide Web
XML	Extensible Markup Language

INTRODUCCIÓN

Debido a la presencia de organismos de acreditación que rigen a las diferentes Instituciones de Educación Superior, las bibliotecas universitarias están cada vez más obligadas a proveer información verificable sobre el uso de los recursos de información que poseen.

La mayor cantidad de bibliotecas, gestiona el registro de sus documentos mediante un SIGB (Sistemas Integrados de Gestión de Bibliotecas), que en su mayoría son programas que gestionan de forma adecuada los datos, pero que comúnmente sólo sirven para el almacenamiento de la información y la elaboración limitada de reportes básicos. El gran problema es cómo hacer para que a partir de la gran cantidad de información que almacenan, está se transforme en conocimiento útil, entendible y al alcance de todos.

El esfuerzo por procesar información ha sido relegado, mayormente, a la tabulación de datos, planteando constantemente desafíos que ayuden a obtener mucha más información, a partir de estos, adoptando estrategias que permitan desarrollar mecanismos para filtrar, seleccionar, desarrollar, analizar e interpretar los datos, con la finalidad de encontrar información oculta que sea útil para la institución.

El uso de la minería de datos es hasta el momento un conjunto de técnicas y tecnologías que permite realizar análisis de grandes cantidades de información, producto de transacciones en BD, pero antes de ser procesada y minada, la información debe sintetizarse para ser útil y en base a ella tomar decisiones para el mejoramiento y desarrollo de las actividades que la empresa quiera mejorar.

Los procesos, y las técnicas utilizadas, son variados al momento de desarrollar un proyecto sobre minería de datos, pero en general, la extracción del conocimiento es una de las más importantes dentro de la minería de datos, pues esta permite la selección, procesamiento y limpieza de los datos, procesos claves en la obtención del conocimiento.

Este trabajo, se enfoca en el desarrollo práctico de un modelo de reglas de asociación, analizando los datos desde esta perspectiva, comenzando con una revisión teórica, mediante una revisión literaria del tema, hasta llevarlo a su aplicación práctica a fin de identificar patrones de consultas sobre el fondo bibliográfico.

Los diferentes procesos aplicados dentro de la herramienta utilizada, Rapidminer, permiten desarrollar desde la carga de información, pasando por el proceso ETL, hasta la implementación del modelo y posterior creación de las reglas de asociación.

CAPÍTULO 1

GENERALIDADES

1.1 Antecedentes

Desde los inicios de la Historia de la Humanidad, los Centros de Información Bibliográfico (CIB), dentro de los cuales se consideran a las Bibliotecas, Centro de Información, Centros de Recursos para el Aprendizaje y la Investigación (CRAI), etc., han sido los guardianes del conocimiento, estos se encuentran estrechamente ligada a la evolución del libro, entendiéndose el libro como contenido. Las Instituciones de

Educación Superior (IES), que poseen CIB, públicas o privadas; como guardianes de la información, se ven obligadas a realizar de forma periódicas evaluaciones sobre el uso de sus colecciones, de tal manera que se puedan identificar posibles falencias o necesidades de información, evaluaciones que se llevan mediante análisis estadísticos, de los usos de sus recursos, con la firme convicción de validar de alguna forma servicios prestados con los recursos económicos asignados.

Dada la importancia que tiene la obtención de datos fiables que permitan medir los servicios, se ven obligadas a implementar mejoras constantes de tecnologías que permita de alguna forma diagnosticar su desempeño, identificar fortalezas y debilidades, tomar acciones en procesos que requieren mejoras, evaluándolos de una forma más o menos precisa, permitiendo incorporar prácticas que ayuden a optimizar los servicios que demandan los usuarios.

Las IES en el Ecuador, atraviesan constantemente por procesos rigurosos para su acreditación, sobre todo desde la entrada en vigor de la LOES y de acuerdo con los procesos de acreditación por parte de las diferentes entidades de control, es así, que resulta imprescindible implementar procesos y procedimientos que mejoren los mecanismos de medición interna, aprovechando que muchos de los datos son procesados de forma semiautomática (en un sistema de información),

usando herramientas ofimáticas como procesadores de palabras u hojas de cálculo, que de alguna manera ayudan en la obtención de información, pero que están aún distantes de hacerlo de forma consolidada, peor aún, de proporcionar un análisis detallado de la información.

1.2 Descripción del problema

Las recientes necesidades de las IES (Instituciones de Educación Superior), específicamente refiriéndonos al campo de los CIB, dedican mucho esfuerzo por contar con datos que ayuden a proporcionar información fiable que permita el estudio de sus usuarios (potenciales y reales), y como estos se relacionan mediante los préstamos realizados con el fondo bibliográfico que poseen, ha impulsado a realizar esfuerzos para la obtención de estadísticas que satisfagan esta necesidad, sin embargo, estas sólo son capaces de brindar parte de la información requerida, pero no brindan una visión perspectiva (qué se está haciendo y cómo se está haciendo) y prospectiva (cómo puede evolucionar la institución) como apoyo a la toma de decisiones.

Generar conocimiento a partir de los datos que se tienen es fundamental, pues, aprovecharlos y potencializarlos permitirá la optimización de presupuestos, renovando o creando nuevos servicios,

así como demostrar que las decisiones relacionadas a las adquisiciones son realizadas de forma estratégica.

En la actualidad, existe dificultad en poder identificar qué tipo de materiales bibliográficos han sido, o no, consultados, evaluar los periodos de alta demanda de información, evaluar qué tipo de información es más requerida según el área de conocimiento, identificar repeticiones sobre las consultas de un determinado tipo de material bibliográfico, establecer tendencias o patrones de consultas dentro del conjunto de datos, identificar grupos afines u homogéneos, los mismos que pueden considerarse, en algún momento, para llevar a cabo investigaciones.

Además, no se puede identificar información de interés para los usuarios, a fin de poder recomendar material bibliográfico utilizado en materias similares, mejorar la difusión selectiva de la información, o identificar la falta de material bibliográfico según la demanda.

Lo explicado anteriormente no permite cumplir a cabalidad con requisitos que establecen determinados criterios de acreditación llevados a cabo por organismos nacionales, como el CEAACES, o recomendaciones que llevan a cabo instituciones internacionales como el IFLA (International Federation of Library Associations and Institutions), en aras de mejorar los servicios relacionados con las

actividades propias de este tipo, como préstamos, **tipos de usuarios que hacen uso de estos recursos**, así como, la pertinencia o el uso adecuado de los fondos para el desarrollo de las colecciones.

1.3 Objetivos

Objetivo General

Implementar una aplicación de minería de datos para la identificación de perfiles de usuarios y patrones de consulta sobre el material bibliográfico de un Centro de Recursos de Información.

Objetivos específicos

- Identificar perfiles de usuarios y patrones de consultas sobre el material bibliográfico.
- Analizar e identificar la pertinencia de los recursos de información.
- Explorar y analizar las asociaciones que existe entre el fondo bibliográfico y las preferencias del usuario.
- Evaluar y difundir los resultados obtenidos.
- Analizar los resultados obtenidos para la toma de decisiones.

1.4 Alcance, restricciones y limitaciones

El conocimiento generado, producto de la investigación, se realizará mediante difusión pasiva de la información (bajo demanda). El usuario interesado solicitará al centro de información, mediante una página Web, y según sus requerimientos; por su parte, el centro de información responderá, sobre la necesidad de información planteada, proporcionando la información por esta misma vía.

Quedan excluidos del análisis el material bibliográfico que no es motivo de préstamos como ediciones incunables, libros raros, obras de referencia y publicaciones periódicas, es decir, material en estantería que por su condición de únicos son reservados y no pueden ser utilizados con alta frecuencia por los usuarios en general. Incluir este tipo de material, que no son representativos, puede alterar, u ofrecer, resultados sesgados.

1.5 Metodología

Debemos entender que la minería de datos es un proceso que involucra un conjunto de técnicas y tecnologías que nos permiten explorar grandes bases de datos de manera automática o semiautomática, en general, la minería de datos intenta ayudar a comprender el contenido de un repositorio de datos, por lo tanto, se recomienda la utilización de un modelo de minería de datos dirigida (model-driven), recomendación

que se estima primordial por ser el inicio del proyecto, debido a que no se cuenta con la maduración necesaria para otro modelo.

1.6 Solución Propuesta

La principal fuente de información de este proyecto será a partir de la toma de datos observacionales, y de los datos existentes en los registros automatizados, es decir, datos productos de las actividades que se llevan a cabo diariamente, específicamente se tomarán datos producto de la circulación y préstamo de material bibliográfico, datos que muestran el nivel de utilización de los diferentes tipos de materiales, y el tiempo que son requeridos, así como otros tipos de materiales bibliográficos similares consultados por un mismo usuario.

Primero, seleccionaremos y cargaremos los datos, es decir, recopilaremos los datos y se definirá qué métodos se utilizará para llevarlo a cabo. Este es un proceso fundamental, del cual dependen los demás pasos, pues, si el conjunto de datos que se van a utilizar no es idóneo, el proceso resultante será erróneo.

Segundo, realizaremos la preparación y el procesamiento de los datos previamente recopilados, disponiendo de un almacén de datos con la información en un formato común y descartando inconsistencias. Se distribuirá los datos, simetría y normalidad de las correlaciones existente en la información.

Tercero, se realizará la interpretación de los resultados, se evaluarán patrones, se revisará la necesidad de nuevas iteraciones, finalmente se difundirá y se hará uso del nuevo conocimiento.

1.6.1 Calendario del Proyecto

A continuación, se presenta el cronograma de acuerdo con el proceso que se va a intervenir.

Tabla 1
Cronograma del proyecto

#	HITO	Mes 1				Mes 2				Mes 3				Mes 4			
		S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
1	Seleccionar el proceso a analizar																
2	Comprender el proceso y su contexto																
3	Lectura del estado del arte																
4	Metodología a utilizar																
5	Proceso ETL																
6	Preparacion y preprocesamiento de datos																
7	Normalización de los datos																
8	Desarrollo de la solución																
9	Evaluación del modelo																
10	Interpretación de los resultados																
11	Retroalimentación y uso del conocimiento																
12	Difusión de los resultados																

1.6.2 Características del Hardware

Tabla 2
Descripción del hardware utilizado

Cant.	Prod.	Características
1	**Computador**	**Laptop** TOSHIBA Satellite L45-B Procesadores Intel Core i5 2.20 GHz cache 3MB Memoria RAM 4 GB. DDR3L 1600MHz Disco Duro 500GB (5400RPM) Pantalla 14.0 pulgadas (1366 x 768) resolución HD Sistema operativo Win 10 Home 64 bits, procesador x64 1 Puertos de salida de video VGA 3 Puertos USB 2.0

CAPÍTULO 2

MARCO TEÓRICO

2.1 Minería de Datos

La minería de datos no aparece por el desarrollo de nuevas tecnologías, sino que se crea, en realidad, por la aparición de nuevas necesidades y, especialmente, por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de información.

La minería de datos es una de las etapas del KDD (Knowledge Discovery in Databases), tal como se aprecia en la figura 2.1, es un

campo de las ciencias computacionales que se refiere al proceso que intenta descubrir patrones en grandes volúmenes en conjunto de datos, utiliza técnicas provenientes de la IA (Inteligencia Artificial) y de la estadística, estas técnicas son algoritmos un tanto complejos, y sofisticados, que se aplican en un conjunto de datos para la obtención de resultados.

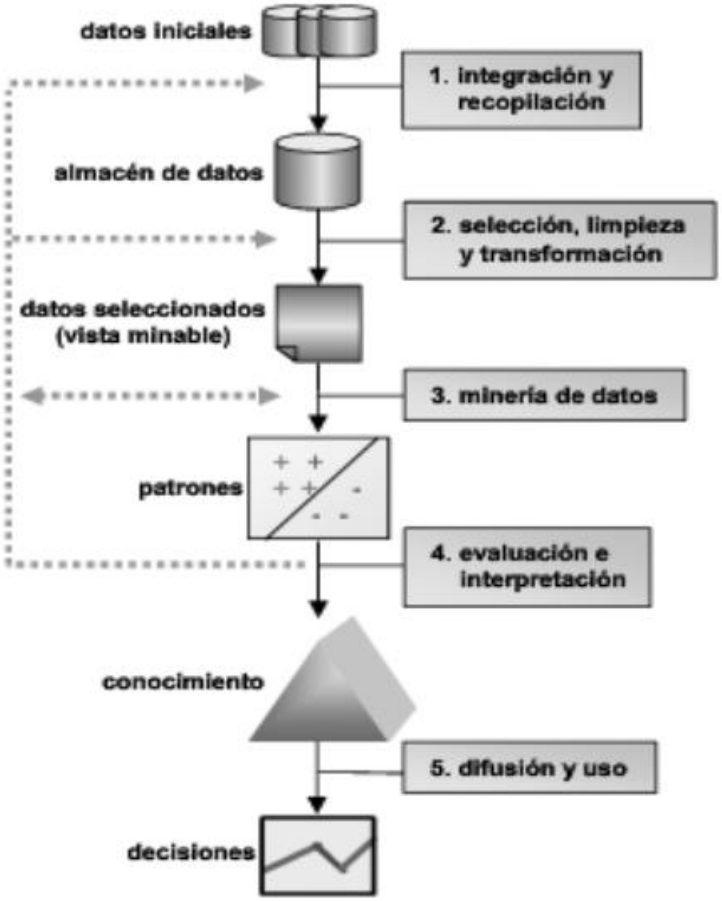


Figura 2.1 : Fase del proceso KDD.
Tomado de “Introducción a la minería de datos” por Hernández Orallo [1]

Para Hernández Orallo, la minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos con la finalidad de encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) y el uso de los patrones debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización [1].

La Minería de Datos comprende cinco fases:

1) **Proceso ETL**. - en esta fase se realiza la extracción, transformación y carga de los datos transaccionales hacia los sistemas de almacenamiento de datos.
2) **Almacén de datos**. - fase en la que se realiza la recopilación y gestión de los datos en un sistema de base de datos multidimensional.
3) **Acceso a los datos**. - definimos el acceso a los datos, es decir, quienes podrán tener acceso a los datos recopilados (analistas de negocios, profesionales TIC, usuarios autorizados.
4) **Análisis de datos**. – es la parte fundamental de la minería de datos, existen diferentes niveles de análisis disponibles, entre los que se destacan modelos gráficos como las redes neuronales artificiales

(aprenden a través de la formación), árboles de decisión (generan reglas para la clasificación de conjuntos de datos), algoritmos genéticos (técnicas de optimización que se basan en combinación genética, mutación y selección natural basada en la evolución natural).

5) **Presentación.** - mediante una tabla, gráfico o algún otro tipo de presentación que permita la correcta interpretación de la información generada.

2.1.1 Tipos de datos

La minería de datos puede ser aplicada a cualquier tipo de información, siendo las técnicas, de minería de datos, diferentes para cada una de ellas. Debemos aprender a diferenciar entre datos estructurados provenientes de bases de datos relacionales, otros tipos de datos estructurados en base de datos (espaciales, temporales, textuales y multimedia) y datos no estructurados provenientes de la web o de otros tipos de repositorios de documentos. Entre los principales tipos de datos tenemos:

Base de datos relacionales. - Se trata de una colección de relaciones, tablas, cada tabla consta de un conjunto de atributos (columnas o campos) y puede contener un gran número de tuplas

(registros o filas), cada tupla representa un objeto, el cual se describe a través de los valores de sus atributos y se caracteriza por poseer una clave única o primaria que lo identifica.

Otros tipos de bases de datos. - a pesar de que las bases de datos relacionales son las más utilizadas, existen aplicaciones que requieren otro tipo de organización de la información, entre las cuales podemos mencionar:

Base de datos espaciales. – los datos que almacenan se encuentran relacionados con el espacio físico (ciudad, región, mapas de posicionamiento global, tráfico, etc.). La minería da datos, permite, por ejemplo, en el caso del transporte, encontrar patrones asociativos relevantes de una zona determinada para la construcción de nuevas líneas de transporte público [2].

Base de datos temporales. – se encargan de almacenar datos históricos, en lugar de datos actuales, a diferencia de las bases de datos instantáneas, las cuales almacenan datos actuales (y que se actualizan cuando los hechos dejan de ser ciertos). Su diseño permite la captura de información que varía en el tiempo, al mismo tiempo que mantiene un historial de los cambios que se producen sobre los datos [3].

Las bases de datos documentales. – es un tipo de base de datos NoSQL, está orientada a documentos, su contenido se basa en las descripciones para los objetos (documentos de texto) que pueden contener desde las palabras clave a los resúmenes.

Estas bases de datos pueden contener documentos no estructurados, semi-estructurados o estructurados, los datos son normalmente almacenados en formato JSON o XML, entre las más conocidas, tenemos a Cassandra, MarkLogic y MongoBD de Amazon [4].

Las bases de datos multimedia. – dentro de sus principales características podemos mencionar el almacenamiento discreto de la información como imágenes, audio y video, objetos de gran tamaño que por lo general se almacenan fuera de las bases de datos (en sistemas de archivos). Este tipo de bases necesita su propia metodología de búsqueda, almacenamiento y sistema de ficheros integrados, junto con métodos de minería de datos.

La indexación de la información es crucial en este tipo de base de datos, sobre todo cuando crecen en volumen. Los objetos multimedia suelen tener atributos como fecha de creación, categoría, creador, con la finalidad de luego almacenar estos datos para su posterior recuperación [5].

WWW. - se refiere a los datos alojados en la web, es el repositorio de información más grande y diverso que existe, desde donde se puede extraer contenido (documentos, hipertextos, imágenes, etc.), estructura (organización de la web, su estructura y como se accede a ella) y conocimiento relevante y útil (extracción de patrones de uso de los usuarios).

Realizar minería en la web no es una tarea sencilla, pues muchos de sus datos no son estructurados o semi-estructurados, la variedad de sus contenidos dificulta poder realizar minería, debido a que estos datos se almacenan en distintos servidores que tienen características particulares [6].

2.1.2 Tipos de modelos

El objetivo principal de la minería de datos es el análisis de datos para extraer conocimiento. Este conocimiento puede estar representado en forma de relaciones, patrones o reglas inferidos de los datos y desconocidos. Estas relaciones, constituyen el modelo de los datos analizados. Hay varias formas de representar los modelos, y cada una de estas, determina el tipo de técnica que puede utilizarse para inferirlos. Ya en la práctica, los modelos pueden ser *predictivos* o *descriptivos* [1].

El **modelo para el análisis predictivo** (aprendizaje supervisado) es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado presente o futuro [7].

Es importante que al momento de realizar un análisis predictivo, se posea una gran cantidad de datos, históricos y actuales, con la finalidad de establecer patrones de comportamiento e inducir conocimiento, este se realiza mediante una serie de procesos que conlleva al aprendizaje computacional, es decir, los computadores estarán en la capacidad de aprender de forma autónoma y de esta forma generar nuevo conocimiento y capacidades [7].

Dentro de las características principales, se encuentra la separación (supone la existencia) de las variables en dos grandes tipos de variables, la variable que se quiere pronosticar (*variable dependiente*) y otro grupo de variables que se utiliza para construir dicho pronóstico (*variables independientes*).

El **modelo descriptivo** (aprendizaje no supervisado), permite cuantificar las relaciones entre los datos, sirve, por ejemplo, para

clasificar clientes o contactos en grupos, identifican las diferentes relaciones que pueden existir entre los clientes y los productos.

Los modelos descriptivos pueden ser de simulación, la teoría de colas o las técnicas de previsión, las observaciones son generalmente clasificadas en grupos que no son conocidos con anterioridad, pudiendo los elementos de las variables estar conectados entre sí, de acuerdo con vínculos desconocidos de antemano, sin la existencia de hipótesis de causalidad, de esta forma, todas las variables disponibles son tratadas al mismo nivel. El análisis descriptivo realiza cálculos estadísticos descriptivos para resumir los datos [7].

En las características del modelo descriptico, podemos mencionar que no hay distinción entre las variables, todas tienen la misma jerarquía (juegan el mismo rol o tiene el mismo estatus), dentro de este modelo tenemos a la Reglas de Asociación, Clustering o Agrupamientos, por lo tanto, nos enfocaremos en las Reglas de Asociación pare el desarrollo del presente trabajo.

2.1.3 OLAP y minería de datos

Las herramientas OLAP (On-Line Analytical Processing), son genéricas, funcionan sobre un sistema de información, permiten combinaciones de datos de manera compleja, proporcionan

facilidades para manejar y transformar los datos, así como la producción de otros datos (agregados o combinados), permitiendo el análisis de los mismos debido a que pueden producir diferentes *vistas.*

Estas herramientas son utilizadas para la toma de decisiones, describen lo que hay en la base de datos, generalmente se utiliza para responder la veracidad de algo, pudiendo aprobarse o rechazarse. En resumen, podemos decir que *OLAP* es un *proceso deductivo.*

Las herramientas de minería de datos son múltiples, permiten la extracción de patrones, modelos, así como, descubrir relaciones, regularidades y tendencias, ayudan a producir reglas o patrones (conocimiento), a diferencia de OLAP, se puede utilizar para analizar una cantidad mayor de variables. La minería de datos difiere con OLAP debido a que, en lugar de verificar patrones hipotéticos, utiliza los mismos datos para el descubrimiento de los mismos patrones, volviéndose un *proceso inductivo* [8,9].

OLAP genera modelos hipotéticos y que mediante queries trata de verificar (proceso deductivo), pero que cuando la cantidad de variables crece, se vuelve una herramienta compleja tomando un tiempo mayor para encontrar la hipótesis. En cambio, la minería

de datos, en lugar de centrarse en modelos hipotéticos (para realizar verificaciones), utiliza los datos para encontrar esos modelos (proceso inductivo), pero en general se tratan de procesos que se complementan entre sí.

2.1.4 Relación de la minería de datos con otras disciplinas

La minería de datos se mueve entre campos multidisciplinares, debido a que por su naturaleza se presente como una extensión de otras disciplinas, entre las principales podemos mencionar:

Estadística. - importante rama del conocimiento, la cual proporciona algoritmos y técnicas que se utilizan en la minería de datos como la media, la varianza, regresión lineal, regresión no lineal.

Matemáticas. – se encuentra relacionados por la utilización de operadores lógicos (cuando se realizan comparaciones para la toma de decisiones) y matemáticos en los procesos de inferencia y construcción de patrones.

Recuperación de información. - recupera la información a partir de datos textuales, el desarrollo se ha basado en el uso efectivo de bibliotecas, así como la búsqueda de Internet, permite localizar documentos a través de las palabras clave.

Base de datos. - se aplican técnicas de indización y de acceso eficiente a los datos que son muy importante para el diseño de algoritmos eficientes de minería de datos. El proceso KDD (del cual es parte el Data Mining) toma los datos que se encuentran normalmente almacenados en las bases de datos.

Inteligencia artificial. - permite el desarrollo de algoritmos para el aprendizaje autónomo del sistema.

Sistema para la toma de decisiones. - mediante herramientas y sistemas informáticos proporcionan información necesaria para la toma de decisiones, tanto en el ámbito empresarial como en el área de la salud.

Visualización de datos. - se refiere a la presentación de los datos, tales como barras, histogramas, gráficos de pastel, etc., con la finalidad de que los especialistas o expertos en cada área puedan visualizar el conocimiento generado [1].

2.1.5 Arquitectura típica de un sistema de minería de datos

Una arquitectura típica de un sistema de minería de datos está compuesta por los niveles que se muestran en la figura 2.2.

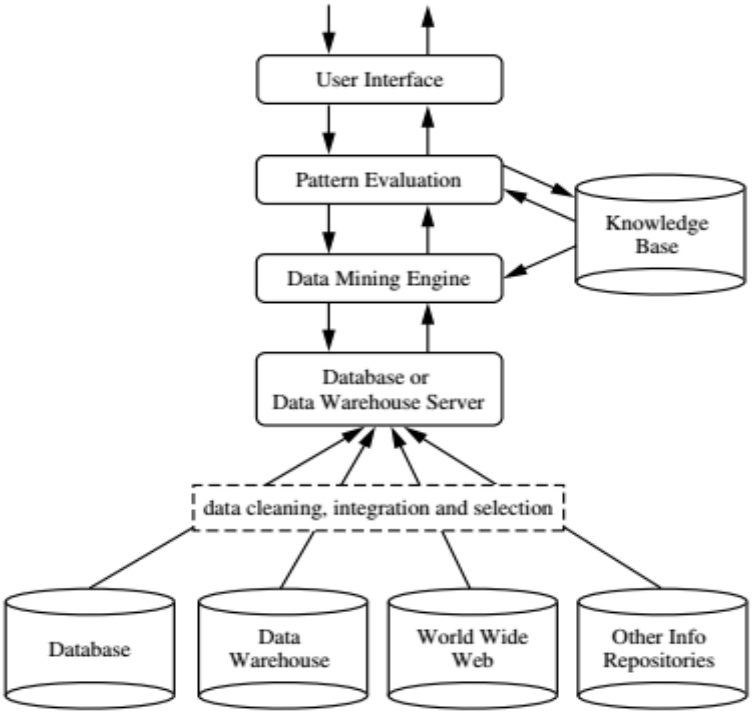


***Figura 2.12** :* Arquitectura típica de un sistema de Minería de Datos. Tomado de “Data Mining: concepts and techniques” por Han y Kamber [10]

Nivel Base, se refiere al almacenamiento de los datos, estos pueden encontrarse dentro de una base de datos, almacén de datos, hoja de cálculo, páginas web u algún tipo de repositorio. Los datos a este nivel, se les aplica técnicas de limpieza e integración.

Base de datos o Datawarehouse server, en esta fase se procede a crear un repositorio unificado para los datos, los mismos que pueden provenir de varios sistemas, se construye de acuerdo con el proceso de minería de datos elegido.

Base del conocimiento, nos ayuda en el conocimiento del dominio para guiar la búsqueda, que se espera de los datos para descubrir comportamientos inesperados, inclusive umbrales de evaluación.

Dentro de este nivel (Base del conocimiento) encontramos la **Ingeniería para la minería de datos**, que consiste en un conjunto de módulos funcionales para las tareas, tales como: la caracterización, asociación y análisis de correlación, clasificación, predicción, análisis de cluster, análisis de valores atípicos y evolución de análisis.

Otro de los factores que integran este nivel (Base del conocimiento) es la **Evaluación de Patrones**, este componente emplea medida de interés e interactúa con los módulos de minería de datos a fin de enfocar la búsqueda hacia patrones interesantes. Se recomienda la evaluación del patrón lo más profundo posible en el proceso de minería para limitar la búsqueda a solo los patrones que interesan.

Interfaz de usuario, este módulo establece la comunicación entre el usuario y el sistema de minería de datos, permitiendo al usuario interactuar con el sistema como por ejemplo la ejecución

de queries o tareas específicas. Permite, además, la evaluación de patrones y su visualización en diferentes formatos [10].

2.1.6 Aspectos estadísticos de la minería de datos

La estadística es el arte y la ciencia de la colección, interpretación y análisis de datos y habilidad de obtener generalidades lógicas relacionadas con un fenómeno bajo investigación, la estadística se encuentra basada en la teoría, enfocada en la prueba de hipótesis, la minería de datos a su vez integra la teoría y heurística (encontrando algoritmos con buenos tiempos de ejecución y buenas soluciones).

Muchos de los conceptos, algoritmos y técnicas (como la media, mediana moda, varianza, análisis univariante y multivariante, regresión lineal y no lineal, técnicas bayesianas, etc.) que se utilizan en minería de datos provienen de la estadística, esta se ocupa de generalizar los resultados obtenidos, a diferencia de la inteligencia artificial que se ocupa de establecer soluciones algorítmicas mediante costos razonables [11].

La minería de datos nace y tiene sus raíces en la inteligencia artificial, mediante el reconocimiento de patrones,

específicamente dentro del Machine Learning[1] y la estadística. Los métodos de minería de datos utilizan conceptos estadísticos como segmentación; por su parte los métodos estadísticos pueden combinarse con técnicas de minería de datos como transformación de datos y reducción de datos.

El profesor e investigador Oldemar Rodríguez nos dice que, a diferencia de la estadística, en la minería de datos no se trabaja con muestras de la información, sino que se buscan patrones ocultos en los datos mediante algoritmos (realmente no se trabaja sobre datos, sino sobre conceptos), utilizando un sistema automatizado [12].

2.1.7 Ética y minería de datos

Mediante el uso que hacemos en Internet, dejamos huellas digitales, mediante el uso de los miles de servicios que se encuentran disponibles, de forma inconscientes permitimos que nuestro perfil sea utilizado para saber, por ejemplo, nuestro comportamiento de compras, qué usualmente nos gusta, dónde

[1] Machine Learning es una rama de la inteligencia artificial que consiste en un método de análisis de datos que realiza de forma automática un modelo analítico usando algoritmos que aprenden de forma iterativa a partir de los datos de interés que se reciben de los usuarios

viajamos, dónde vivimos, y una serie de interrogantes vinculadas a los hábitos que ejecutamos diariamente.

La información obtenida por la Minería de Datos puede utilizarse para discriminar a las personas, y visto desde este punto de vista, esto no es ético, y de cierta forma ilegal. Las compañías involucradas en el levantamiento de la información, debe tener políticas claras que permitan que el usuario siempre sepa cómo y donde serán utilizados sus datos. Se ha dado casos en los que muchos de los datos recopilados para otro propósito terminan en manos de personas o empresas poco escrupulosas con el uso de la información.

Muchas de las investigaciones sociales pueden verse afectada por la mala manipulación de los datos, por ejemplo, Facebook tuvo una filtración masiva (más de 50 millones de usuarios) de los datos de sus usuarios, los mismos que fueron utilizados con fines políticos por parte de la empresa Cambridge Analytica y su matriz, Strategic Communication Laboratories, las cuales realizaron segmentación de mensajes electorales (en la campaña del entonces presidenciable estadounidense Donald Trump) para influenciar en los potenciales votantes.

Debemos comprender, como en el caso de Facebook, que ante la falta de integridad y ética en las empresas, lo que constituyen un peligro en la manipulación de los datos, los académicos también, son tentados a filtrar los datos que aparentemente son utilizados con fines académicos.

En el caso de Facebook, los datos en un principio eran utilizados para las actividades académicas, Aleksandr Kogan (investigador de la Universidad de Cambridge) desarrolló una aplicación para Fecebook, la cual recopilaba datos de los usuarios y de sus contactos (con fines académicos), que al final terminaron en manos de la empresa Cambridge Analytica [13].

A nivel internacional el **Código Internacional ICC/ESOMAR** para la práctica de la investigación de Mercados, Opinión y Social y del Análisis de Datos menciona:

Art 1. Deber de cuidado

a) Los investigadores deben asegurarse de que los titulares de los datos (personas o instituciones) no se vean perjudicados como consecuencia directa del uso de sus datos personales en una investigación.

Art 5. Uso de los datos secundarios

Cuando se empleen datos secundarios que incluyan datos personales los investigadores deben asegurarse de que:

a) El uso pretendido es compatible con el propósito para el que originalmente se recogieron los datos.
b) Los datos no se recogieron violando las restricciones legales, mediante engaño, o de manera que no era aparente o razonablemente discernible o prevista por el titular de los datos.

Además, dentro de sus principios fundamentales consta que los investigadores deben ser transparentes en relación a la información que se proponen recoger, así como, el llevar una conducta ética, no llevando a cabo nada que pudiera causar daño al titular de los datos, o perjudicar la reputación de la investigación [14].

Por su parte la **UNESCO**, en su Conferencia General, llevada a cabo en París en el año 2011, en el anexo “Código de Ética para la Sociedad de la Información”, manifiesta en el literal 12 que “Toda persona tiene derecho a la protección de sus datos personales y su vida privada en Internet y otras TIC. Se debe proteger a los usuarios contra la conservación ilícita y la

divulgación indebida o no autorizada de dichos datos, y contra la intromisión en su vida privada" [15].

En Europa, la **Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal**, así como la **Directiva 95/46/CE del Parlamento Europeo y del Consejo**, establece que "todo individuo es titular de los datos que le conciernen y le afectan personalmente", pero el principal problema es que en su mayoría, estos datos no están bajo la custodia de la persona que los genera, por el contrario, la custodia de esta información son las empresas que prestan algún tipo de servicio en la red, quienes tienen la capacidad tecnológica para recopilarlos y explotarlos, es decir, son los verdaderos dueños de la información [16].

A nivel local, el **Código Orgánico de la Economía Social de los Conocimientos** (conocido también como Código Ingenio), menciona, en el Capítulo V, aspectos a tener presente la Ética en la Investigación Científica:

Art. 67.- Ética en la investigación científica. - Los principios necesarios para el cumplimiento de la ética en la investigación científica estarán desarrollados en un Código Ético Nacional, el cual deberá contemplar al menos los siguientes ámbitos:

1. El respeto por la dignidad de la vida y la biodiversidad;

2. Consentimiento informado de las personas partícipes en investigación;
3. Consentimiento previo, libre e informado de pueblos y nacionalidades;
4. Respeto y protección de los derechos de las personas partícipes en investigación;
5. Confidencialidad de los datos personales, así como aquellos exceptuados en el Código Ético Nacional, obtenidos en procesos de investigación; y,
6. Respeto a los animales con fines de experimentación [17].

En resumen, los investigadores y las personas involucradas en la manipulación de los datos deben cumplir con normas y criterios que permitan el buen hacer de la profesión, respetando los derechos que el usuario tiene sobre sus datos, llevando a cabo la investigación mediante consentimiento informado o guardando la privacidad o confidencialidad que corresponde a fin de evitar aspectos legales negativa.

2.2 Técnicas de minería de datos

Dentro del proceso KDD, debemos tener claro la diferencia que existe entre las "***tareas***" y las "***técnicas***" (métodos o algoritmos) empleadas

para resolver dicha tarea, es así, que las tareas pueden ser predictivas y descriptivas.

Las *tareas predictivas* se refieren a la minería supervisada, mientras que, las *tareas descriptivas* incluyen aspectos no supervisados.

Por otro lado, la técnica elegida dependerá mucho de proyecto a desarrollar, los objetivos que este persigue y de la calidad, cantidad y las características propias de los datos. Sobre este tema, hablaremos con más detalles en la sección “Técnicas de minerías evaluadas”.

2.2.1 Tareas predictivas

Clasificación. - es una de las técnicas más utilizadas, cada instancia, o registro de la base de datos, pertenece a una clase (clasifica un dato dentro de una de las clases categóricas predefinidas). Permite emparejar o asociar datos a grupos predefinidos mediante aprendizaje supervisado, encuentra modelos que describen y distinguen clase o conceptos para futuras predicciones.

Regresión. - consiste en aprender una función real que asigna a cada instancia un valor real, es decir, permite corresponder un dato con un valor real de una variable. Una de las principales diferencias con la técnica de clasificación, es que, el valor a

predecir es numérico, tiene como objetivo, minimizar el error entre el valor predicho y el valor real.

2.2.2 Tareas descriptivas

Clustering (agrupamiento). - es una de las técnicas más utilizadas, permite la agrupación de características preestablecidas de acuerdo a criterios de distancias o similitud, es un método no supervisado. Un cluster es una colección de registros que son similares entre sí, pero distintos a los que pertenecen a otro cluster.

Dentro de las subcategorías, destaca el Clustering de Documentos, el cual realiza descubrimientos sobre agrupaciones de documentos, tomando como base contenidos o mediante un determinado tema que caracteriza a cada uno de los grupos en los que se logra dividir cada grupo, este clustering, requiere una correcta selección de características y del algoritmo a utilizar.

En otra subcategoría, encontramos al Clustering Difuso, el cual se ha utilizado satisfactoriamente en la Minería de Texto, por su adecuada gestión de datos incompletos o que generan “ruido” en la información, mejorando la comprensión de los patrones y para el tratamiento de los clusters que se superponen.

Reglas de asociación. - se utilizan para descubrir hechos que ocurren de forma común dentro de un determinado conjunto de datos (causa – consecuencia), pueden descubrir irregularidades en transacciones registradas en grandes repositorios de datos.

Para que exista una consecuencia, se debe considerar cada posible combinación de condiciones. Las reglas de asociación cumplen un rol importante en la cobertura o soporte (número de instancias predichas de forma correcta) y la precisión o confianza (proporción de número de instancias que es aplicada a la regla) [18].

2.2.3 Técnicas de minerías evaluadas

A continuación, describiremos alguna de las técnicas más utilizadas dentro de la minería de datos y del aprendizaje automático:

Árboles de decisión. - Es una técnica de aprendizaje supervisada, donde existen regiones no superpuestas y hojas, en la cual deben conocerse los resultados esperados con la finalidad de identificar las decisiones y reglas necesarias para llegar a ellos, es una técnica utilizada, sobre todo, para realizar predicciones.

Son parte del modelo predictivo, permite la construcción de diagramas de forma lógica a partir de la información que contiene una base de datos, se utiliza en la minería de datos con la finalidad de tomar decisiones convenientes desde el punto de vista probabilístico.

Una de las principales características de esta técnica, una vez que han sido analizados los datos, es su utilización para la solución de problema de predicción, clasificación y segmentación. Dentro de los algoritmos que utiliza tenemos a el Algoritmo ID3, C4.5 (evolución del ID3) y el algoritmo CART (el cual no es de decisión, pero se lo utiliza para problemas de regresión).

Dentro de su representación gráfica consta de nodos de probabilidad, nodos de decisión, nodos terminales y ramas; el primero se representa con un circulo y muestra las probabilidades de ciertos resultados; el segundo se representa con un cuadrado y muestra una decisión que se tomará; y un nodo terminal nos muestra el resultado definitivo de una ruta de decisión; por su parte, la rama nos indica los distintos caminos a seguir cuando se toma una decisión o ante el suceso de un evento aleatorio.

A través de un ejemplo vamos a ilustrar de una mejor forma su proceso (ver figura 2.3), donde tenemos un producto X, y se desea dentro de un grupo de personas con un rango de ingresos y edades, quienes tienden a comprar más ese producto.

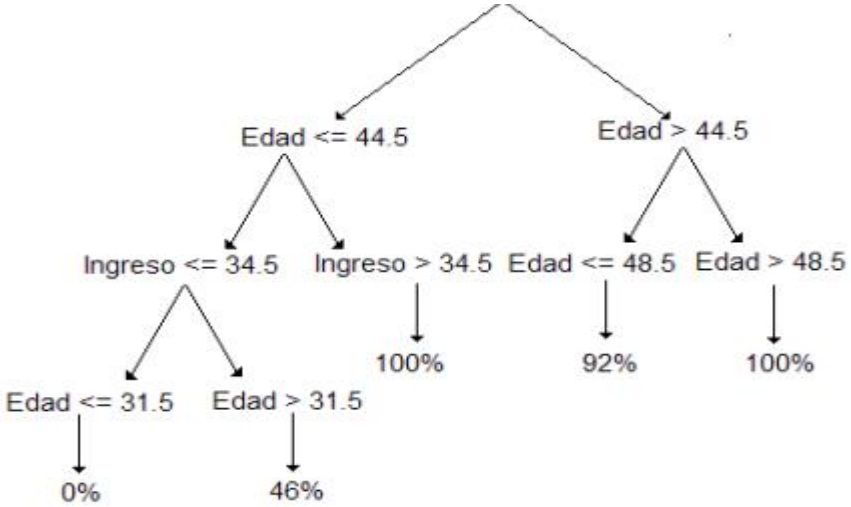


Figura 2.2 : Árbol de decisión para el producto X.
Tomado de “Minería de Datos” por Ferruccio et al. [19]

Redes Neuronales. - es una técnica mayormente utilizada en problemas de detección de patrones y detección de características comunes en los datos, sirve para el tratamiento de datos complejos, de los cuales tenemos escaso conocimiento, una de las principales ventajas de utilizar esta técnica, es que no realizan suposiciones sobre la naturaleza de la distribución de los datos, trabajan bien con los datos incompletos u ocultos.

Dentro de sus principales características, podemos mencionar que tienen la capacidad de aprender y generar conocimientos a partir de datos incompletos, mediante entrenamiento, mejorando una función en particular ajustando los valores de las conexiones entre los elementos (ver figura 2.4). Dentro de los principales tipos de redes neuronales, de acuerdo al tipo de aprendizaje, podemos mencionar el Supervisado, No supervisado y por Corrección, su aplicabilidad dependerá de las necesidades y objetivos del proyecto.

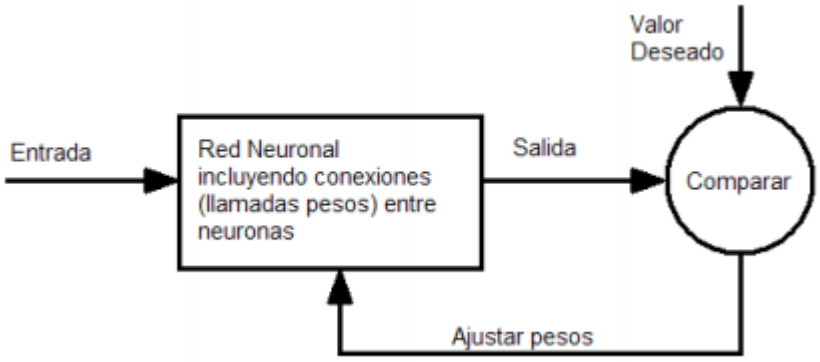


Figura 2.4 : Entrenamiento de una Red Neuronal.
Tomado de “Minería de Datos” por Ferruccio et al. [19]

Para las redes neuronales supervisadas y por Corrección, se requiere un grupo de datos de control con los que se debe verificar los resultados obtenidos con la finalidad de generar conocimiento. Por el contrario, las técnicas no supervisadas, no requieren datos de control para verificar los resultados obtenidos,

esto porque se consideran todos los datos de entrada como variables aleatorias, generando conocimiento [16][17].

Algoritmos genéticos. - los algoritmos genéticos imitan la evolución de los seres vivos mediante la mutación., reproducción y selección., estos funcionan a partir de una población inicial de elementos (cromosomas), generados aleatoriamente, luego, se ejecutan las funciones de variación y selección hasta alcanzar un criterio establecido, por sus características, son ampliamente utilizados en problemas que requieren optimización de procesos de algoritmos.

Utilizan mecanismos como: sobrevivencia de los organismos con mejor capacidad dentro de una población, el uso de secuencias de caracteres (uno y ceros) en string como representación de ADN de dichos organismos y el uso de métodos aleatorios para la generación de la población y su reproducción.

Redes Bayesianas. - permiten el aprendizaje sobre las relaciones de dependencia y casualidad, combinan conocimiento con datos, evitando el sobre ajuste de los datos, además, del manejo de bases de datos incompletas. Son una representación gráfica de dependencias para razonamiento probabilístico, dentro de la cual, los nodos representan variables aleatorias y los

arcos representan relaciones de dependencia directas entre las variables

La obtención de una red bayesiana a partir de los datos es un proceso de aprendizaje, el cual contiene dos aspectos: aprendizaje paramétrico, mediante una estructura, obtiene las probabilidades a priori y las condiciones requeridas, y aprendizaje estructural, la cual obtiene las relaciones de dependencia e independencia entre las variables involucradas.

K-NN (K Nearest Neighbours). - este algoritmo cada dato nuevo en el grupo que corresponde, según K vecinos más cerda de otro o de un grupo. Se puede utilizar en clasificación y regresión, en el caso de clasificaciones utiliza el histórico que poseemos. Cuando se quiere realizar un pronóstico sobre una nueva observación, basamos este pronóstico no en todas las observaciones disponibles, sino, en las observaciones que se parecen más a l nueva observación.

Opera a través de la distancia euclídea, y se calcula mediante la fórmula:

$$\textbf{(2.2.1)}\ d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

A priori .- fue propuesto en el año 1994 por Agrawal y Srikant, dado un umbral C, este algoritmo identifica todos los conjuntos de ítems que son subconjunto de al menos C transacciones en la base de datos, permitiendo encontrar de forma eficiente "*conjunto de ítems frecuentes*", los que a su vez sirven para generar reglas de asociación.

En su primera fase, encuentra los Itemset frecuentes, es decir, los ítems que van juntos y que superan la medida de soporte, luego, en su segunda fase, toma los ítems que superan la medida de soporte y a partir de allí considera las reglas que superan la medida de confianza.

Uno de los principales problemas es la extracción de las reglas a partir de los datos proporcionados, debido a que tiene un costo (computacional) muy alto, volviendo imposible la generación de dichas reglas, por ejemplo, para una empresa dedicada a las ventas, la cual puede poseer miles de ítems (2^n -1).

Debido a la naturaleza de nuestro proyecto, trabajaremos con algoritmos relacionados con reglas de asociación, específicamente FP-Growth, el mismo que una variante mejorada del algoritmo Apriori.

2.3 Sistemas y herramientas para la minería de datos

La selección de la herramienta idónea, para la minería de datos, permitirá realizar de la mejor forma el proceso de extracción del conocimiento, para realizar la respectiva selección, debemos tener en consideración:

- **El tipo de dato a analizar**. - no todos los datos son iguales, así como, las herramientas tampoco están listas para el tratamiento de todos los tipos de datos, debiéndose comprobar los formatos con los que trabaja la herramienta seleccionada.
- **Compatibilidad entre sistemas**. - debe ser compatible con el sistema sobre el cual va a operar, debemos comprobar si existe interfaces de usuarios basadas en la web, así como la manipulación de los datos de entrada en formato XML.
- **Funciones y metodologías**. - debemos saber si la herramienta seleccionada posee una única función o es capaz de proporcionar múltiples funciones (como la descripción, análisis estadístico, clasificación, predicción, agrupamiento, etc.). Se debe tener muy claro las funciones que vamos a utilizar de la herramienta elegida.
- **Acoplamiento con BD o SAD** (Sistemas de Almacenamiento de Datos). - deben acoplarse con los sistemas de almacenamiento

de datos o de bases de datos, propiciar el procesamiento desde un entorno uniforme, pudiéndose producir sin acoplamiento, acoplamiento débil, acoplamiento semi-estanco y acoplamiento estanco.

- **Escalabilidad**. - puede estar constituida por la escalabilidad de fila (el sistema se considera escalable cuando el número de filas se amplía en una determinada proporción y el sistema no tarda más tiempo del estimado para ejecutar la consulta) o escalabilidad de columna (el tiempo de ejecución de consulta aumenta linealmente a la vez que lo hace el número de columnas).
- **Herramientas de visualización**. - se debe tener en cuenta la visibilidad que ofrecerá, en función de los datos y resultados o mostrar todo el proceso.
- **Interfaz gráfica de usuario**. - evaluar la interacción que la herramienta posee para la interacción y la fácil presentación de los resultados.[22]

Considerando las características anteriormente expuestas, existen algunas aplicaciones que se ajustan a las características que deben poseer las herramientas diseñadas para la extracción del conocimiento, entre las que se destacan:

IBM SPSS Modeler. – originalmente llamada SPSS Clementine, y que a partir del 2009 se pasó a llamar PASW Modeler, luego, con la adquisición por parte de IBM, se pasó a llamar IBM SPSS Modeler, herramienta que se centra en la integración de Data Mining con otros procesos y sistemas de negocio que proporcionan inteligencia predictiva de forma eficiente durante las operaciones del negocio.

Versiones recientes incluyen funcionalidades como reglas de scoring y modelos de árboles de decisión, permitiendo encontrar patrones, dentro de la información, a fin de facilitar la toma de decisiones.

Su interfaz gráfica simple pone, el poder que representa la minería de datos, en manos de los usuarios de negocios y el alto rendimientos aumenta la productividad de los analistas.

RapidMiner (YALE). – es una herramienta creada por la universidad de Dortmund, es de código abierto con licencia pública GPL de GNU[2], es multiplataforma, debido a que está hecho en JAVA; flexible para el descubrimiento del conocimiento y la Minería de Datos en un ambiente

[2] La licencia GPL de GNU, se usa para la mayoría de los programas GNU y en más de la mitad de los paquetes de software libre, está orientada principalmente, a proteger la libre distribución, modificación y uso de software, con el propósito de declarar que el software cubierto por esta licencia es libre y protegerlo contra intentos de apropiación que restrinjan esas libertades a los usuarios.

de experimentos en aprendizaje automático, utilizado en tareas de minería de datos, así como en el mundo real.

En sus últimas versiones acepta nuevos formatos de entrada de datos (Microsoft Excel y SPSS), ofrece representación de datos en formato 2D y 3D, así como representación en formato SOM (Self Organizing Map). Ofrece más de 500 operadores para los principales procedimientos de máquinas de aprendizaje.

Weka. – Es una herramienta creada por colaboradores de la Universidad de Waikato, contiene licencia GPL para su libre distribución, contiene una gran cantidad de algoritmos que trabajan de forma eficiente, posee, además, una gran cantidad de herramientas para realizar una gran cantidad de tareas para Minería de Datos.

Mediante programación en JAVA puede programarse algoritmos sofisticados para análisis de datos y creación de modelado predicativo. Posee una interfaz gráfica intuitiva para el acceso a sus funcionalidades, se pueden implementar técnicas de clasificación, asociación, agrupamiento y predicción, además, es un software multiplataforma.

Orange. – es una suite de software para minería de base de datos y aprendizaje automático basado en componentes que cuenta con un fácil y potente, rápido y versátil front-end de programación visual para el

análisis exploratorio de datos y visualización, y librerías para Python y secuencias de comando. Tiene complementos para el pre-procesamiento de datos, características de puntuación y filtrado, modelado, evaluación del modelo y técnicas de exploración.

Knime (Konstanz Information Miner). – plataforma de código abierto, con una curva rápida de aprendizaje para la integración, procesamiento, análisis y exploración, permite al usuario crear, de forma visual, flujos o tuberías de datos, ejecutar parte o todos los pasos del análisis para, posteriormente, estudiar los resultados, modelos vistas interactivas.

Está desarrollado en Java y basado en Eclipse, permite el uso de plugin para la extensión de funcionalidades [23,24]. La tabla 3, nos muestra, en resumen, las características que poses cada una de las herramientas tomadas en consideración para el proyecto, la cual se seleccionará de acuerdo al trabajo que se pretende realizar.

Tabla 3
Herramientas de Minería de Datos

Nombre	Tipo/Lic.	Plataforma	Arq.	Algoritmos
PASW Modeler (Clementine SPSS)	Propietaria	Multiplataforma	C/S	Clustering, regresión, redes neuronales, árboles de decisión

Nombre	Tipo/Lic.	Plataforma	Arq.	Algoritmos
RapidMiner	Freeware/ versiones de pago	Multiplataforma	C/S	Clustering, árboles de decisión, redes neuronales, reglas de asociación
Weka	Free/ GPL	Multiplataforma	C/S	Clustering, regresión, árboles de decisión, redes neuronales
Orange	Free/GPL	Multiplataforma	C/S	Clustering, árboles de clasificación, reglas de asociación
Knime	Free/GPL	Multiplataforma	C/S	Algoritmos de segmentación, árboles de decisión, redes neuronales

Nota: C/S = Cliente Servidor

En general los distintos tipos de softwares utilizados para realizar Minería de Datos, analizan relaciones y patrones entre los datos generados, utilizando técnicas estadísticas, redes neuronales y aprendizaje automático, que permiten detectar clases (para la localización de datos en grupos predeterminados), clusters (para identificar afinidades de consumo o realizar segmentación de mercados), asociaciones (para identificar tendencias o hábitos) y patrones secuenciales (para revelar comportamientos encadenados).

Por lo tanto, la herramienta elegida dependerá específicamente del proyecto que se pretenda implementar, de las características particulares que presenten los datos, y lo más importante, la necesidad de información que requieran los interesados.

CAPÍTULO 3

DEFINICIÓN DE LA SITUACIÓN ACTUAL

3.1 Diagnóstico de la situación actual

Un Centro de Recursos Bibliográficos, es una dependencia que funciona en un entorno dinámico, en el mismo, se integran los servicios de información que posee una universidad, entre las variedades en cuanto a su nombre, podemos encontrar a los CRAI (Centros de Recursos para el Aprendizaje y la Investigación), Redes de Bibliotecas o Centros de Información, en general responden a contener

información, y se encuentran relacionadas a la tecnologías que brindan soporte al aprendizaje, la docencia y la investigación.

Los servicios que se prestan en el Centro de Recursos de Información son utilizados por usuarios que demandan información para la realización de tareas o investigaciones, entre los servicios que podemos mencionar se encuentran: préstamos de libros impresos, préstamos de libros digitales, prestamos de tesis impresas y digitales, préstamos de equipos de apoyo para actividades académicas o de investigación, préstamos de cubículos y salas de estudios, talleres para la formación de investigadores, entre los más destacados.

Para la prestación de estos servicios se debe contar con personal de carácter multidisciplinario como bibliotecarios, informáticos y técnicos para brindar un soporte adecuado a las necesidades del usuario (estudiantes, profesor, investigador, etc.).

Los servicios mencionados con anterioridad generan una serie de registros, es así como cada día, en promedio, se reciben entre 1.500 a 1.700 visitas, se generan 500 préstamos de material impreso, 300 préstamos de equipos informáticos y 50 reservas de espacios de trabajo.

La tabla 4 muestra las cifras correspondientes a los préstamos generados por los principales servicios que se ofrecieron durante los años 2016 y 2017, además, la tabla 5 presenta los 10 libros más solicitados durante

los periodos antes mencionados, estos corresponden a aquellos que se encuentran presentes acompañados con otros ítems en una transacción.

Tabla 4
Datos estadísticos de los servicios

Servicio del Centro de Recursos de Información	Año 2016	Año 2017
# de visitas	266.043	223.818
# de préstamos de material bibliográfico (impreso)	20.503	37.381
# de préstamos de material bibliográfico (digital)	3.070	1.515
# de búsquedas en BD de información científicas	395.299	953.477
# de búsquedas en el Repositorio Digital	121.220	555.462

Tabla 5
Libros más solicitados

#	Título	# prestamos	Fracción
1	Gaceta Judicial	1.025	0,030
2	Harrison: principios de medicina interna	884	0,026
3	Anatomía humana	856	0,025
4	Atlas de anatomía humana	551	0,016
5	Periodontología clínica e implantología odontológica	537	0,013
6	Odontología restauradora: fundamentos y técnicas	455	0,010
7	Tratado de anatomía humana	322	0,008
8	Libro básico de ortodoncia	269	0,008

#	Título	# prestamos	Fracción
9	Cecil y Goldman: tratado de medicina interna	262	0,007
10	Gran atlas McMinn de anatomía humana	244	0,007

Nota: libros más solicitados en transacciones que contienen más de un ítem

3.2 Características e interpretación de la fuente de datos

Los datos, objeto del análisis en el presente trabajo, recogen información de los préstamos de material bibliográfico realizado por las diferentes carreras (27 carreras) que se imparten en la institución, abarcan distintos aspectos descriptivos de los usuarios y del personal que interviene en el préstamo (de acuerdo con el servicio prestado).

Entre las variables más importantes destacamos: el usuario responsable del préstamo (usuario administrativo), fecha de préstamo, tipo de literatura, numero secuencial, inventario, tema del material consultado, autor del material consultado, tipo de usuario (estudiante, profesor, investigador, etc.), ciclo de estudio, identificación, tipo de préstamo (interno o domicilio), estado del préstamo (fecha de devolución o en circulación), carrera a la que pertenece y nombre del usuario.

Los datos que se analizarán serán a partir de reportes generados en formato textos, la población no se encuentra definida, se cuenta con archivos históricos en base a documentos proporcionados para la investigación. El diseño y la solución del modelo, beneficiará de forma

directa, al Centro de Recursos Bibliográfico, y de forma indirecta a la institución.

3.3 Definición de requerimientos y restricciones

- Se debe contar con los datos a analizar, los mismos deben estar en formato texto.
- No se modificarán la tecnología existente, el sistema a evaluar deberá leer la información que se entregará en formato texto.
- Describir las técnicas y métodos que se han utilizado para la recogida de datos.
- Por confidencialidad de la información, no se mencionarán algunos datos reales como la información de actores o datos técnicos.
- El software elegido debe ser seguro en todas las fases del proceso, no se debe compartir información no autorizada.
- Construir un prototipo visual (sistema) para la aplicación de Minería de Datos.
- Analizar los resultados obtenidos para la aplicación de la Minería de Datos.
- Descubrir información relativa a la asociación que existen entre los ítems en determinadas transacciones.
- Emitir las recomendaciones y predicciones para el empleo de la Minería de Datos en un Centro de Recursos de Información.

3.4 Levantamiento de la información

3.4.1 Descripción de los procesos

En este apartado proporcionaremos de forma detallada los procesos relacionados con los préstamos de material bibliográfico, actividades que permiten la generación de los datos. A continuación, describiremos los procesos que contribuyen en la generación de los datos que serán explorados:

Tabla 6
Proceso de préstamo de material bibliográfico impreso

Pasos	Departamento	Descripción
1	Atención a usuarios	Recepta el requerimiento del usuario (información requerida).
2	Atención a usuarios	Verifica el requerimiento dentro sistema (existencia de información requerida).
3	Atención a usuarios	Solicita el documento para el préstamo de material bibliográfico
4	Atención a usuarios	Registra e imprime el préstamo de material bibliográfico al usuario
5	Atención a usuarios	Recuerda las políticas de préstamo (fecha de devolución, sanciones, etc.).
6	Atención a usuarios	Desactiva mecanismos de seguridad y entrega el material bibliográfico solicitado.

Tabla 7

Proceso de devolución del material bibliográfico impreso

Pasos	Departamento	Descripción
1	Atención a usuarios	Recepta el material bibliográfico
2	Atención a usuarios	Verifica fecha de devolución
3	Atención a usuarios	Si ha pasado la fecha de devolución, el sistema imprime recibo de multa
4	Atención a usuarios	Explica a usuario el motivo de la multa generada
5	Atención a usuarios	Firma el recibo y solicita al usuario la firma correspondiente
6	Atención a usuarios	Entrega recibo y activa la seguridad en el material bibliográfico
7	Atención a usuarios	Devuelve el material bibliográfico a la percha correspondiente

Tabla 8

Proceso de préstamo de material bibliográfico digital

Pasos	Departamento	Descripción
1	Sistema Bibliotecario	Usuario realiza la búsqueda en el OPAC

Pasos	Departamento	Descripción
2	Sistema Bibliotecario	Verifica información del material bibliográfico requerido
3	Sistema Bibliotecario	Ingresa a la plataforma de préstamo, con usuario y clave respectiva
4	Sistema Bibliotecario	Busca dentro de la plataforma seleccionada el material requerido
5	Sistema Bibliotecario	Verifica disponibilidad (en ciertos casos) y demás detalles del material bibliográfico
6	Sistema Bibliotecario	Descarga, o lee en línea, el material bibliográfico seleccionado
7	Sistema Bibliotecario	Devolución del material bibliográfico

3.4.2 Identificación de la fuente de conocimiento

En esta sección, describiremos los medios y tecnologías utilizadas en el Centro de Recursos de Información, así como, las técnicas y métodos utilizados para la toma de datos.

Describir y analizar los procesos, relacionados con los datos, existentes y definir los procesos, usuarios y servicios utilizados es fundamental para la obtención de los datos y posterior procesamiento de la información.

Los datos serán obtenidos a partir del sistema utilizado, mediante un archivo de texto, el cual contiene los datos principales de las transacciones realizadas a diario por los usuarios.

La recopilación de la información se realizó de la siguiente forma:

a) Identificación y selección del Centro de Recurso de Información participante, existen varios centros de recursos de información en toda la institución, las cuales generan transacciones sobre un sistema centralizado.
b) Análisis de la información, contiene información sobre el Centro de Recursos de Información seleccionado, y los principales campos referentes a la información que se requiere analizar. Los datos corresponden a los periodos 2016 y 2017.

3.4.3 Descripción de la fuente de información

Antes de llevar a cabo un estudio estadístico de la información, es importante conocer los datos recopilador y como estos colaborarán en el estudio, de la misma forma, debemos describir en que variables se encuentran organizado. Detallaremos las variables cuantitativas y cualitativas y se describirá las características principales y la vinculación que posee con los datos.

Variables Cualitativas. - indican cualidades o características que no pueden ser medidas de forma numérica, se divide a su vez en variables nominales (o categóricas) y ordinales (guardan un orden), las primeras tienen como característica, que sus datos tienen eventos mutuamente excluyentes y colectivamente exhaustivos, a diferencia de las variables ordinales, las cuales se caracterizan por que sus variables guardan un orden dentro de las categorías.

Tabla 9
Operacionalización de las variables cualitativas

Nombre de la variable	Tipo	Escala de medición
Tema	Cualitativa	Nominal
Autor	Cualitativa	Nominal
Ciclo	Cualitativa	Nominal
Usuario responsable	Cualitativa	Nominal
Tipo de usuario	Cualitativa	Nominal
Editorial	Cualitativa	Nominal
Año de edición	Cualitativa	Nominal
Idioma	Cualitativa	Nominal
Tipo de literatura	Cualitativa	Nominal
Sexo	Cualitativa	Nominal

Variables Cuantitativas. - representan características que pueden ser representadas de forma numérica, de tal forma que pueden

realizarse operaciones numéricas sobre ellas. Existen 2 subcategorías dentro de este tipo de variables: la variable cuantitativa discreta, la cual toma un *número finito* de valores entre dos valores cualesquiera de una característica y la variable cuantitativa continua, en la que se puede tomar un *número infinito* de valores entre dos valores cualesquiera de una característica.

Tabla 10
Operacionalización de las variables cuantitativas

Nombre de la variable	Tipo	Escala de medición
Transacción	Cuantitativa	Ordinal
Fecha de préstamo	Cuantitativa	De razón (días)
Fecha de devolución	Cuantitativa	De razón (días)

3.4.4 Objetos del negocio

A continuación, se detallan los principales objetos del negocio que contribuyen con la generación de los datos.

Tabla 11

Descripción de los objetos del negocio

Nombre	Tipo	Descripción	Parámetros	Roles involucrados
Usuario (cliente)	BO	Representa información personal del usuario que solicita el material bibliográfico.	**Usuario** • Nombre. • Apellido. • Facultad. • Ciclo. • Correo. • Teléfono. • Dirección.	Cliente.
Material Bibliográfico	BO	Representa el registro del material bibliográfico.	**Material Bibliográfico** • # Inventario • Clasificación • Presentación • Año • Editorial • Edición	Procesos técnicos.
Documento de identidad	BO	Representa el documento que habilita el préstamo del material bibliográfico	**Documento** • Tipo • # Identificación • Nombre • Apellidos • Vigencia	Usuario. Auxiliar de Biblioteca.
Solicitud de préstamo de material bibliográfico impreso	BO	Representa la solicitud del préstamo del material bibliográfico.	**Solicitud** • # Transacción • Tipo de préstamo • Tipo de usuario • Código de usuario • Código de material • Fecha de préstamo. • Fecha de devolución.	Usuario. Auxiliar de Biblioteca.
Registro para devolución del material bibliográfico	BO	Representa la devolución del material bibliográfico	**Registro:** • Código del material bibliográfico • Fecha de devolución.	Usuario. Auxiliar de Biblioteca
Registro para préstamo del material bibliográfico digital	BO	Representa el préstamo del material bibliográfico digital	**Registro:** • Id de plataforma • Id de usuario	Usuario Sistema Bibliotecario

Nombre	Tipo	Descripción	Parámetros	Roles involucrados
Orden de pago	BO	Representa la orden para el pago por multa	**Orden de pago:** • Fecha de orden pago. • Valor de pagar.	Auxiliar de Biblioteca. Usuario.
Solicitud de estadísticas de uso de material bibliográfico	BO	Representa la solicitud de información sobre el uso del fondo bibliográfico	**Reporte:** • Fecha de inicio • Fecha de fin • Tipo de usuario	Coordinador del Centro de Recursos Bibliográfico

Nota: la columna tipo representa el objeto de negocio **(**Business Object)

3.4.5 Matriz de Casos de Usos

Tabla 12

Detalle de la matriz de casos de uso del proceso analizado

ID	Actividad	Tipo	Descripción	Rol	Objeto de Negocio	Posibles estados finales
A1	Solicitud del material bibliográfico.	Manual	El usuario llena los datos solicitando el material bibliográfico.	Usuario	Código de clasificación del material Autor Tema	
A2	Verificar la solicitud del usuario	Manual	La auxiliar de biblioteca verifica si hay disponibilidad de la información.	Auxiliar de biblioteca	Código de clasificación de material Autor Tema	
A3	Ofrecer alternativas de información	Manual	La auxiliar de biblioteca ofrece opción de información ante la falta del material solicitado.	Auxiliar de biblioteca	Código de clasificación de material Autor Tema	
A4	Verificar documentos para el préstamo	Manual	La auxiliar de biblioteca verifica el documento entregado por el usuario	Auxiliar de biblioteca	Vigencia de documento Documento válido para préstamo	

ID	Actividad	Tipo	Descripción	Rol	Objeto de Negocio	Posibles estados finales
A5	Generación del préstamo de material bibliográfico	Manual	La procede al registro del material bibliográfico al usuario	Auxiliar de biblioteca	Fecha de devolución Datos del usuario Datos del material	
A6	Preparar el material a entregar	Manual	La auxiliar procede a desactivar las seguridades del material bibliográfico	Auxiliar de biblioteca	Fecha de devolución	
A7	Entrega del material bibliográfico	Manual	La auxiliar entrega el material bibliográfico y recuerda las políticas de préstamo y devolución	Auxiliar de biblioteca	Fecha de devolución	
A8	Renovación del material bibliográfico	Manual	La auxiliar procede a atender y registrar la fecha de renovación.	Auxiliar de biblioteca	• Fecha de renovación.	
A9	Reporte de estadísticas de uso de material bibliográfico	Manual	El coordinador procede a recopilar la información, luego la procesa para mostrar los datos requeridos	Coordinador del Centro de Recursos de Información	• Fecha de inicio • Fecha fin • Tipo de usuario • Carrera	

Nota: la columna ID representa las actividades

3.4.6 Excepciones

Tabla 13
Registro de excepciones del proceso analizado

ID	Excepción	Actividad Afectada	Descripción	Acciones Correctivas	Objeto de Negocio
E1	Material bibliográfico No disponible	A2	El material no está disponible para el usuario.	El personal debe comunicarse con el usuario y ofrecer otro material.	• Inventario. • Tipo de material.

ID	Excepción	Actividad Afectada	Descripción	Acciones Correctivas	Objeto de Negocio
E2	Documento no válido	A4	El documento presentado no es válido para el préstamo.	El auxiliar de biblioteca explica la situación al usuario, procede a descartar la solicitud.	• Documento de Identidad.
E3	Deudas pendientes	A4	El usuario mantiene deudas pendientes con la institución, no se le puede prestar material a domicilio	El auxiliar de biblioteca comunica al usuario que no puede proceder con el préstamo a domicilio debido a que mantiene deudas con la institución.	• No. de orden de pago. • Fecha máxima de pago
E4	NO Renovación del material bibliográfico	A7	El material bibliográfico solicitado no puede ser renovado debido a que cumplió el máximo de renovaciones	El auxiliar de biblioteca explica al usuario los motivos por los cuales no puede realizar la renovación del material bibliográfico.	• Fecha de renovación. • Registro de novedades

Nota: la columna ID representa la codificación en relación a las excepciones

3.5 Identificación de actores y responsabilidades

3.5.1 Registro de información de los Actores Clave

Tabla 14

Actores claves en la generación de los datos

Nombre	Cargo	Dpto.	Dirección	Teléfono	Email
Juan Mejía	Coord./Director del Centro de Recursos	Biblioteca General	Av. Carlos Julio Arosemena	25555555	juanm@ie.educa.com.ec
Ana Drouet	Jefa de Procesos Técnicos	Biblioteca General / Procesos técnicos	Av. Carlos Julio Arosemena	25555555	anad@ie.educa.com.ec

Nombre	Cargo	Dpto.	Dirección	Teléfono	Email
María Gómez	Auxiliar de Biblioteca	Biblioteca General / Atención a usuarios	Av. Carlos Julio Arosemena	25555555	mariog@ie.educa.com.ec
Ruth Alvarado	Auxiliar de Biblioteca	Biblioteca General / Atención a usuarios	Av. Carlos Julio Arosemena	25555555	falvarado@ie.educa.com.ec
Ana Solís	Auxiliar de investigación	Biblioteca General / Área virtual	Av. Carlos Julio Arosemena	25004040	estelas@ie.educa.com.ec
Raúl Gómez	Auxiliar de soporte TICs	Biblioteca General / Área virtual Área virtual	Av. Carlos Julio Arosemena	25004040	raulg@ie.educa.com.ec

3.5.2 Actores y relación con el proceso

Los actores, así como las responsabilidades, que intervienen en el proceso de préstamo de material bibliográfico se muestran a continuación:

Tabla 15

Actores y su relación con el proceso

Actores	Rol que juega	Descripción	Interés	Responsabilidad
Estudiantes, profesores, investigadores y público en general.	Usuarios (Clientes)	Es la persona que solicita algún material bibliográfico.	Solicitar un texto que cumpla con sus necesidades de información	Solicitar el tipo de material bibliográfico. Presentar el documento de identificación. Firmar el documento del registro de préstamos.

Actores	Rol que juega	Descripción	Interés	Responsabilidad
Juan Mejía	Director del centro de información	Es la persona que coordina los servicios que se prestan al usuario	Poner al servicio de los usuarios el mejor material bibliográfico, a fin de que estos puedan contar con información fiable para el desarrollo de sus tareas o investigaciones.	Gestiona la compra o convenio del material bibliográfico impreso o digital. Coordina el uso adecuado de los recursos de información. Ayuda en el diseño de los entornos de acceso a los recursos de información.
Ana Pérez	Jefa de procesos técnicos	Es la persona que se encarga del ingreso, registro, codificación y procesamiento del material bibliográfico	Registrar el material bibliográfico para que esté disponible a la brevedad posible.	Clasificación y catalogación del material bibliográfico en el sistema bibliotecario.
María Gómez; Ruth Alvarado	Auxiliar de Biblioteca	Es la persona que se encarga del proceso de préstamo de material bibliográfico impreso.	Es la imagen del departamento para la atención de los usuarios.	Revisar disponibilidad de material solicitado. Revisar documento de identificación entregado. Registrar el préstamo y devolución de material bibliográfico.
Ana Solís	Auxiliar de investigación	Es la persona que se encarga del préstamo de los equipos para uso académico.	Garantizar la accesibilidad del usuario hacia los recursos de información.	Asignar los equipos para el préstamo de material bibliográfico. Orientar al usuario con el uso de las herramientas y el portal elegido.

Actores	Rol que juega	Descripción	Interés	Responsabilidad
Raúl Gómez	Auxiliar de soporte TICs	Es la persona que se encarga de la capacitación de los usuarios con el entorno virtual.	Capacitar y mejorar el uso de los recursos de información.	Dictar charlas y talleres para el uso de los recursos de información. Mantener en condiciones operables los equipos.

3.6 Descripción de las relaciones entre elementos

A continuación, mediante el modelo Entidad-Relación (ver figura 3.1), podemos ver parte de la estructura de la BD y una descripción de las tablas que intervienen en la generación de los datos sujetos de análisis en el proyecto.

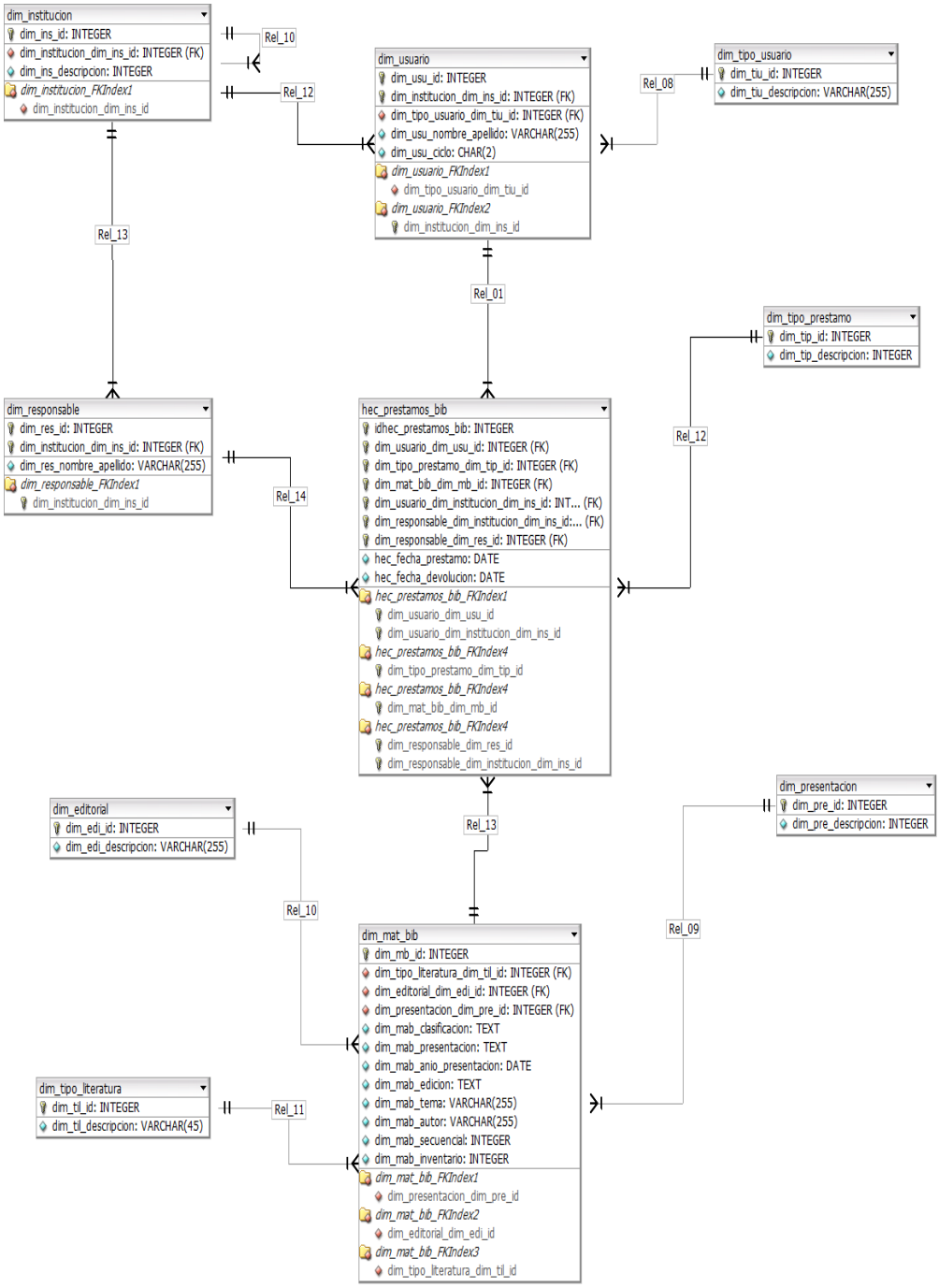


Figura 3.1 : Modelo Entidad Relación

CAPÍTULO 4

ANÁLISIS Y DISEÑO DE LA APLICACIÓN DE MINERÍA DE DATOS

4.1 Análisis, exploración y preparación de los datos

4.1.1 Descripción de las variables objetivas

Describiremos los datos obtenidos con la finalidad de brindar una visión más clara de las variables seleccionadas. Cada transacción representa uno o varios ítems solicitados.

Tabla 16
Descripción de los datos

Variable	Descripción	Minable
Usuario Responsable	Usuario que realiza el préstamo del material bibliográfico	NO
Fecha de préstamo	Fecha en la cual se realiza el préstamo al estudiante. Formato DD-MM-YYYY	NO
Tipo de préstamo	Categoriza al tipo de préstamo (interno o domicilio)	NO
Tipo de literatura	Describe al tipo de material bibliográfico	NO
Num. Secuencial	Número del registro del material bibliográfico en la base de datos	NO
Inventario	Inventario que le corresponde al ítem en la base de datos	NO
Transacción	Corresponde a la transacción realizada por un determinado usuario	SI
Tema	Tema o título del material bibliográfico	SI
Autor	Autor(es) del material bibliográfico	SI
Área del conocimiento	Área del conocimiento a la cual pertenece la obra	NO
Tipo de usuario	Clasificación del tipo de usuario que realiza el préstamo (Estudiante, profesor, investigador, etc.)	NO
Ciclo	Ciclo de estudio al cual pertenece el usuario (si es estudiante)	NO
Num. Identificación	Código de identificación del estudiante	NO

Variable	Descripción	Minable
Estado Devolución	Registra el estado del préstamo del material bibliográfico (fecha de devolución o si se encuentra en circulación).	NO
Carrera/Dpto.	Carrera, especialidad, unidad académica o departamento al que pertenece el usuario	NO
Nombre	Nombre del usuario que realiza el préstamo	NO

4.1.2 Vista minable conceptual

La vista minable, en general, hace referencia al conjunto de datos más relevantes dentro del contenedor de datos, con la idea de agruparlos de tal forma que se puedan procesar con facilidad. Este modelo se realiza luego de las etapas de limpieza y construcción de los datos a analizar. Por su lado, la vista minable conceptual (VMC), describe de forma detallada cada una de las variables que pueden ser consideradas para llevar a cabo la tarea de minería de datos.

Tabla 17
Usuario – Vista minable conceptual

Variable	Descripción	Minable
Tipo de usuario	Estudiante, Profesor, Investigador, Usuario externo, Administrativo, Funcionario	Categorización del usuario

Variable	Descripción	Minable
Ciclo	Ciclo académico al cual pertenece el estudiante	Mediante el ciclo, se puede agrupar los préstamos realizados por los usuarios, se pueden utilizar para recomendaciones

Tabla 18

Préstamo – Vista minable conceptual

Variable	Descripción	Minable
Transacción	Número secuencial que registra la transacción (préstamo)	Sirve como identificador de las transacciones que se realizan, se agrupará para obtener los ítems relacionados.
Fecha de préstamo	Fecha en la que se realiza el préstamo	Servirá para establecer tendencias de las fechas con mayor rotación de material bibliográfico
Tipo de préstamo	Interno Domicilio	Permitirá conocer qué tipo de préstamos tiene mayor demanda

Tabla 19

Material bibliográfico – Vista minable conceptual

Variable	Descripción	Minable
Código secuencial	Código del registro al cual pertenece la obra dentro del sistema bibliotecario	Determinará el código al cual pertenece ese ítem, el cual demostrará que tan solicitado puede ser un título en particular
Inventario	Número secuencial (inventario) del material bibliográfico	Determinará la rotación de un ítem en particular
Tipo de literatura	• Libro • Tesis • Revista	Permitirá identificar qué tipo de material bibliográfico tiene una mayor demanda
Título	Describe el tema de la obra	Se podrá verificar los temas relacionados en cada préstamo, se considerará las transacciones con más de dos items
Autor	Describe al autor de la obra	Puede permitir buscar relaciones entre los autores mediante los títulos de los libros
Área del conocimiento	Área del conocimiento con la cual está relacionada la obra.	Ayudará a reconocer que áreas del conocimiento tiene mayor demanda de material bibliográfico

Tabla 20
Unidad académica – Vista minable conceptual

Variable	Descripción	Minable
Institución	Unidad académica	Identificará la unidad académica, servirá para conocer que unidad realiza mayores prestamos

4.1.3 Vista minable operativa

La vista minable operativa (VOM por sus siglas en inglés), se refiere a la carga e integración de los datos (integrados en una vista), detallados en la vista minable conceptual.

A continuación, describiremos las diferentes variables, así como, la fuente de almacenamiento de donde provienen (ver tabla 21).

Tabla 21
Vista minable operativa

Variable	Descripción	Fuente
fec_prestamo	Fecha de préstamo	Sistema
tip_prestamo	Forma en la que se realiza el préstamo (interno o domicilio)	Sistema
tip_literatura	Especificación del material bibliográfico (Libro, tesis, revistas)	Sistema

Variable	Descripción	Fuente
num_secuencial	Número del registro del material bibliográfico en la base de datos	Sistema
num_inventario	Inventario que le corresponde al ítem en la base de datos	Sistema
transacción	Número de la transacción	Sistema
tema	Tema o título del material bibliográfico	Sistema
autor	Autor(es) del material bibliográfico	Sistema
are_conocimiento	Área a la que pertenece el material bibliográfico (CCSS, CCMM, Filosofía, etc).	Sistema
tip_usuario	Clasificación del tipo de usuario que realiza el préstamo (Estudiante, profesor, investigador, etc.)	Sistema
institucion	Unidad académica	Sistema

4.2 Representación, transformación y clasificación de los datos

Mediante un procedimiento de ETL, se obtendrá los datos que intervendrán en las operaciones de la minería de datos, se seguirá los siguientes pasos:

a) **Selección**. - los datos principales que se analizaran se encuentran en un archivo txt, en el mismo se encuentran los datos generados por los préstamos de material bibliográfico del año 2016 y 2017.

Estos datos son obtenidos a partir del Sistema Bibliotecario que actualmente se utiliza.

Se cargarán el archivo txt a una hoja de Excel, mediante la opción de importación de archivos, y se revisará posibles inconsistencias en la información. Este archivo es que finalmente leerá la herramienta seleccionada.

b) **Pre-procesamiento/Limpieza**. - esta etapa realizaremos labores de limpieza (completar datos faltantes, suavizar el ruido, identificar datos redundantes, valores perdidos, etc.), integración de múltiples fuentes de datos (de existir), transformación (normalización), reducción (reducción del volumen) y discretización (transformación de variables numéricas en categóricas).

c) **Carga**. - cargar de los datos al algoritmo elegido, para el proceso de minería de datos, luego de lo cual se creará las respectivas reglas de asociación.

4.3 Elección de herramientas para computación de los datos

El arte de tomar buenas decisiones permite la obtención de buenos resultados, no podemos elegir la misma herramienta para resolver diferentes problemas, cada problema tienes situaciones particulares y se desarrolla en sus propios escenarios, debemos entender el contexto, datos, resultados y la información relacionada, por lo tanto, es necesario

saber evaluar de forma correcta las herramientas para la solución y toma de decisiones.

En el siguiente trabajo se utilizará, preferentemente, herramientas open source, específicamente para el trabajo de minería de datos, utilizaremos RapidMiner, una aplicación que posee un enfoque visual para representar las tareas de data mining, proporcionando una manera amigable para el tratamiento de los datos, herramienta que es respaldada ampliamente por quienes frecuentemente realizan minería de datos, esto lo demuestra una encuesta realizada (sobre unos 3.000 participantes) en el año 2014 por la empresa KDnuggets, tal como lo muestra la figura 4.1.

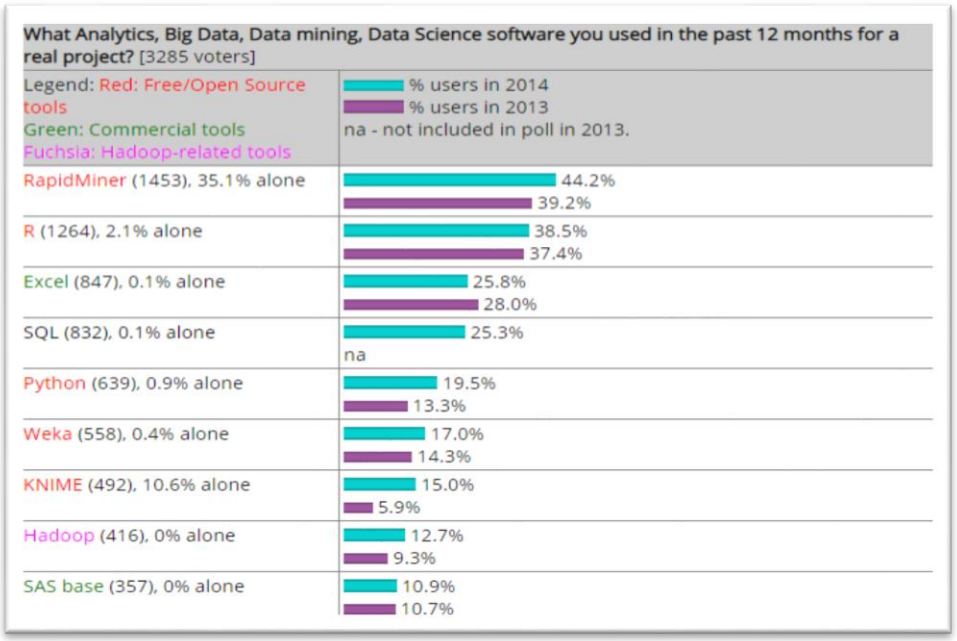


Figura 4.1 : Encuesta sobre usos de software para minería de datos. Tomado de https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html por Piatetsky y KDnuggets [25]

Dentro de las características funcionales de RapidMiner, podemos mencionar que, cada operación se representa mediante un ícono, u operador, los mismos que son conectados mediante una secuencia lógica que representa los pasos y operaciones a los cuales serán sometidos nuestros datos.

4.4 Diseño de la aplicación

4.4.1 Diseño del almacén de datos

Los almacenes de datos son sistemas que almacenan datos históricos, que pueden ser utilizados por herramientas de Minería de Datos, son de consultas y están enfocados a la extracción de conocimiento, debemos tener presente que un almacén de datos no es una base de datos donde se puede realizar funciones básicas de procedimientos almacenados, también denominados CRUD o IMUD por sus siglas en inglés (Insert, Modify, Update o Delete).

Un almacén de datos se encuentra organizados en relación a los temas que más relevancia tiene para la organización, y de alguna forma deben responder a la interrogantes planteadas, proporcionando respuestas estratégicas, distinguiéndose aspectos como las “Actividades” de interés para el análisis y el “Contexto” de análisis para estas actividades de interés,

creándose un modelo multidimensional el cual basa su estructura en ***Hechos*** (se encuentran relacionadas las actividades de interés para la empresa) y ***Dimensiones*** (son el contexto sobre el que se desea analizar las actividades).

La creación del almacén de datos, en nuestro proyecto, se realizará mediante operadores que posee la herramienta y que permiten realizar una serie de procesos. En la herramienta utilizada vamos a segmentar los procesos (utilizando los operadores) que compondrán el ETL (Extrantion, Transformation and Loading), fundamental dentro de la creación de la arquitectura del almacén de datos.

4.4.2 Identificación y detección de datos anormales

Dentro de las acciones que debemos considerar ante la presencia de *datos anómalos* (outliers) tenemos:

Ignorar: muchos de los algoritmos son robustos ante la presencia de datos anómalos.

Filtrar: Mediante la eliminación o reemplazo de la columna con datos anómalos, teniendo cuidado que al eliminar una de las columnas, puede existir columnas dependientes con una mayor

calidad. Mediante el reemplazo podemos agregar una nueva columna indicando si el valor era normal o outlier.

Filtrar la fila: existe un sesgo en los datos, debido a que la mayor parte de las veces las causas de un dato erróneo se encuentran relacionadas con caos o tipos especiales.

Reemplaza el valor: por un valor nulo si el algoritmo lo trata bien o por máximos y mínimo, o por medias. En muchas ocasiones se puede predecir a partir de los otros datos presentes.

Discretizar: al transformar un valor continuo en uno discreto (por ejemplo, muy alto, alto, medio, bajo, muy bajo), permite que los outliers caigan en ‘muy alto’ o ‘muy bajo’ sin mayores problemas.

Por otro lado, para los *datos faltantes* (missing values), debemos considerar las siguientes acciones:

Ignorar: algunos algoritmos son robustos ante la presencia de datos faltantes.

Filtrar: mediante la eliminación o reemplazo de la columna, teniendo en consideración que pueden existir columnas dependientes con datos de mayor calidad. Es preferible reemplazar la columna con otra que contenga datos booleanos, indicando la existencia, o no, de dicho valor.

Filtrar la fila: se evidencia sesgo de los datos, en muchas ocasiones las causas de un dato faltante están relacionadas con casos o tipos especiales.

Reemplazar el valor: por medias, en muchas ocasiones se puede predecir a partir de otros datos.

Segmentar: se segmentan las tuplas por los valores que tienen disponibles, obteniéndose modelos diferentes para cada segmento, los cuales luego se combinan.

Modificar la política: sobre la calidad de los datos, esperando hasta que los datos faltantes estén disponibles.

Normalmente son estos procesos los que permiten extraer los datos de las fuentes de datos transaccionales, realizan las transformaciones necesarias para cargarlos en el almacén de datos, así también es importante mencionar que debe realizarse las cargas sucesivas o refrescos de los datos de acuerdo a las necesidades del negocio [26].

En nuestro proyecto, se revisará y corregirá, de ser necesario, los datos faltantes y anómalos utilizando los operadores que proporciona la herramienta, dentro del proceso ETL (ver figura 4.2).

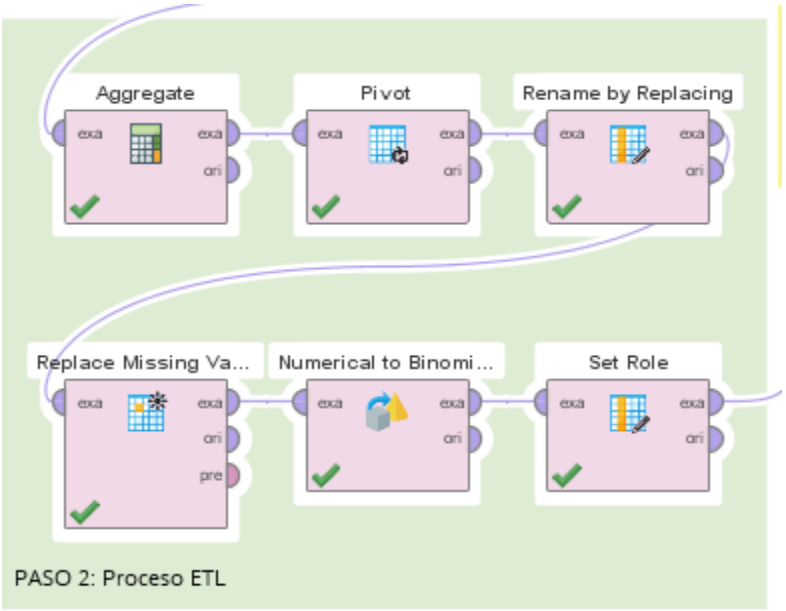


Figura 4.2 : Fase del proceso ETL del proyecto

4.4.3 Validación del modelo

En el presente trabajo, dentro del proceso de minería de datos, se utilizará el modelo de reglas de asociación, se tomará las transacciones que contengan más de un ítem (transacciones en las cuales se solicitaron más de un ítem o libro), se pretende demostrar los vínculos (asociaciones) que existen entre los ítems en determinadas transacciones realizadas por los usuarios.

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Al momento se han investigado diversos métodos para aprendizaje de reglas de asociación, los mismos que han

resultado de mucho interés para descubrir las relaciones, en grandes conjuntos de datos, que se presentan entre variables.

El modelo de minería de datos basado en reglas de asociación se define de la siguiente forma:

- Sea $\mathbf{I = \{s_1, s_2, \ldots s_n\}}$ un conjunto de N atributos binarios llamados *Ítems,* en nuestro caso corresponde al número secuencial asignado a un título de libro.
- Sea $\mathbf{T = \{t_1, t_2, \ldots. t_n\}}$ un conjunto de transacciones (almacenadas en una base de datos correspondientes a los libros requeridos por un usuario en un préstamo), donde cada transacción $\mathbf{t_i}$, es un conjunto de ítems tal que $\mathbf{t_i \subseteq I, 1 \leq i \leq n}$.

Cada transacción en **T** tiene un ID (identificador) único y contiene un subconjunto de ítems de **I**. Una regla se define como una implicación de la forma:

$\mathbf{X \rightarrow Y,}$ es una Regla de asociación donde $\mathbf{X, Y \subseteq I,}$ $X \cap Y = \emptyset$ y $\mathbf{X \cup Y \subseteq t_i}$, los conjuntos **X** y **Y** son mutuamente excluyentes, $\mathbf{t_i}$ es el conjunto de ítems formado por aquellos que corresponden al antecedente o al consecuente de la regla de asociación. El conjunto $\mathbf{X \cup Y}$, debe estar contenido o ser igual a alguna de las transacciones perteneciente a **T**.

Los conjuntos de ítems **X** se denominan "*Antecedentes*" (parte izquierda) y los conjuntos de **Y** "*Consecuentes*" (parte derecha) de la regla.

Para tener una mejor comprensión del tema, Pinho [27] nos muestra mediante un ejemplo (ver tabla 22) , un conjunto de transacciones constituidas por un conjunto de ítems.

Tabla 22
Conjunto de transacciones

ID Transacción	Conjunto de ítems
T1	{a, b, c}
T2	{a, d, b, c}
T3	{a, e, c}
T4	{d, b}
T5	{a, b, e}
T6	{d, b}
T7	{a, d, b, c}
T8	{c, d}
T9	{a, d, b}
T10	{a, b, d}

Mediante la transacción 1 de la tabla 22, podemos generar las siguientes reglas de asociación (ver tabla 23).

Tabla 23
Reglas generadas a partir de la transacción 1

Regla	Conjunto de ítems
R1	a → b, c
R2	a → c, b
R3	a → b
R4	a → c
R5	b → a, c
R6	b → c, a

Regla	Conjunto de ítems
R7	b → a
R8	b → c
R9	c → a, b
R10	c → b, a
R11	c → a
R12	c → b

Las reglas de asociación tienen el objetivo de encontrar tendencias que puedan ser utilizadas para comprender y explorar patrones de comportamiento en los datos, pero, no todas las reglas de asociación representan un patrón en los datos, pues una regla representará un patrón solo si la misma cumple determinados *criterios,* o *medidas de interés*, definidos en los algoritmos de inducción, los cuales también expresan la fiabilidad de las reglas, por lo tanto, para obtener reglas de asociación válidas, debemos definir algunas medidas de interés [27].

4.4.4 Medidas de interés

Uno de los principales problemas cuando procesamos reglas de asociación es que podemos encontrar muchas, o demasiadas, reglas, por lo tanto, hay que limitar el número de reglas para volver manejable el procesamiento posterior, esto nos obliga a analizar las siguientes medidas de interés:

Soporte

Sea una regla **X** → **Y**, el soporte de esta regla se define como el número de veces o la frecuencia (relativa) con que **X** y **Y** aparecen juntas (**X**∪**Y**) en una base de datos transaccional, por lo tanto, definiremos al soporte como:

(4.4.1) **soporte de (X) =** $\frac{|X|}{|T|}$ **(para ítems individuales)**

(4.4.2) **soporte de (X → Y) = P (X∩Y) / T (para ítems de la regla, donde T representa al número de transacciones)**

El soporte puede definirse para ítems individuales, pero también, para la regla. Uno de los primeros requisitos que podemos imponer para limitar el número de reglas, es que tenga un soporte mínimo.

Basándonos en la regla número cuatro, ***a*** → ***c***, del ejemplo descrito en la tabla 23, podemos concluir que el soporte de esa regla es igual al 40% (4/10=0.40), debido a que los ítems ***a*** y ***c*** aparecen cuatro veces (en las transacciones 1, 2, 3 y 7 de la tabla 22) juntos dentro de las 10 transacciones.

Confianza

Sea una regla **X** → **Y**, la confianza de esta regla es el cociente del soporte de la regla y el soporte del antecedente solamente, para lo cual tenemos que:

(4.4.3) Confianza de (X → Y) = Soporte de (X → Y) / Soporte de (X)

Si analizamos el ejemplo, de la tabla 23, de la regla número cuatro, tenemos que la confianza de ***a*** → ***c*** es 0,57 o 57%, puesto que la probabilidad de ***a*** → ***c*** es 0.40 y la probabilidad de ***a*** es 0.70. Esto quiere decir que, si ***a*** se encuentra en una transacción, entonces existe la probabilidad del 57% de que ***c*** esté también en esa transacción.

Debemos diferenciar que el soporte mide la frecuencia, por su parte la confianza mide la fortaleza de la regla.

Lift

Sea una regla **X** → **Y, lift** (sustentación en español), se define de la siguiente forma:

(4.4.4) Lift (X → Y) = Soporte de (X → Y) / { Soporte de (X) * Soporte (Y) }

Lift es una medida que se utiliza para evaluar el grado de dependencia de los términos de una regla. Por ejemplo, en una regla del tipo (**X → Y),** el **Lift** representa el grado en qué **Y** tiende a ser frecuente cuando **X** ocurre, o viceversa. La evaluación de una regla de asociación puede ser realizada de la siguiente forma:

- Si **Lift** (**X → Y) = 1**, o está cerca de 1, indica que la relación es producto del azar.
- Si **Lift** (**X → Y) > 1**, la ocurrencia de los ítems de **Y** influye en la probabilidad de la ocurrencia de los ítems de **X**.
- Si **Lift** (**X → Y) < 1**, la ocurrencia de los ítems de **Y** influye en la probabilidad de la no ocurrencia de los ítems de **X**, por lo que, se deben descartar las reglas que presenten un lift menor a 1.

Si analizamos la regla número 4 de nuestro ejemplo (*a → c*) de la tabla 23, se puede apreciar que **Lift** (*a → c*) = soporte de (*a → c*) / soporte de a * soporte de *c* = (0.40) / (0.70)(0.50) = 1.1428, lo que significa que existe alta probabilidad de que los ítems ***a*** y ***c*** ocurriesen juntos [27,28].

Otras definiciones que considerar

Items: cualquier objeto (producto, paciente, evento) que sea parte de una transacción.

Itemset: una colección de uno o más ítems que se pueden encontrar en una o más transacciones.

Frequent Itemset: un elemento cuyo soporte es mayor o igual a un umbral mínimo.

El proceso para la creación del modelo, así como, las respectivas reglas de asociación, la realizaremos mediante el uso de operadores como *"FP-Growth"* para el algoritmo y *"Create Association"* para la creación de las reglas de asociación, operadores que se encuentran disponibles en la herramienta utilizada (ver figura 4.3).

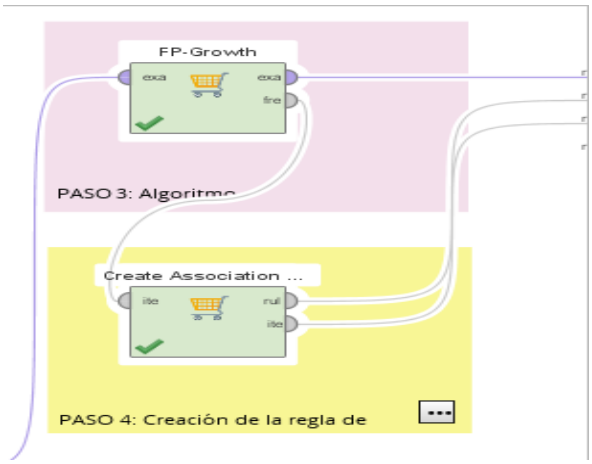


Figura 4.3 *:* Selección de los operadores para la creación del modelo

CAPÍTULO 5

IMPLEMENTACIÓN DE LA APLICACIÓN

5.1 Análisis de las relaciones entre variables

Primero debemos crear un nuevo proceso o espacio de trabajo, luego debemos cargar los datos, esto lo realizaremos mediante la opción “Import Data”, los cuales guardaremos en la carpeta “datos”, desde ahí los cargaremos arrastrándolos hacia el espacio de trabajo, al ejecutarlo, produce el resultado que se observa en la figura 5.1.

Vamos a fijarnos en las columnas, autor, título y transacción, estas son las variables que utilizaremos para desarrollar el proceso ETL.

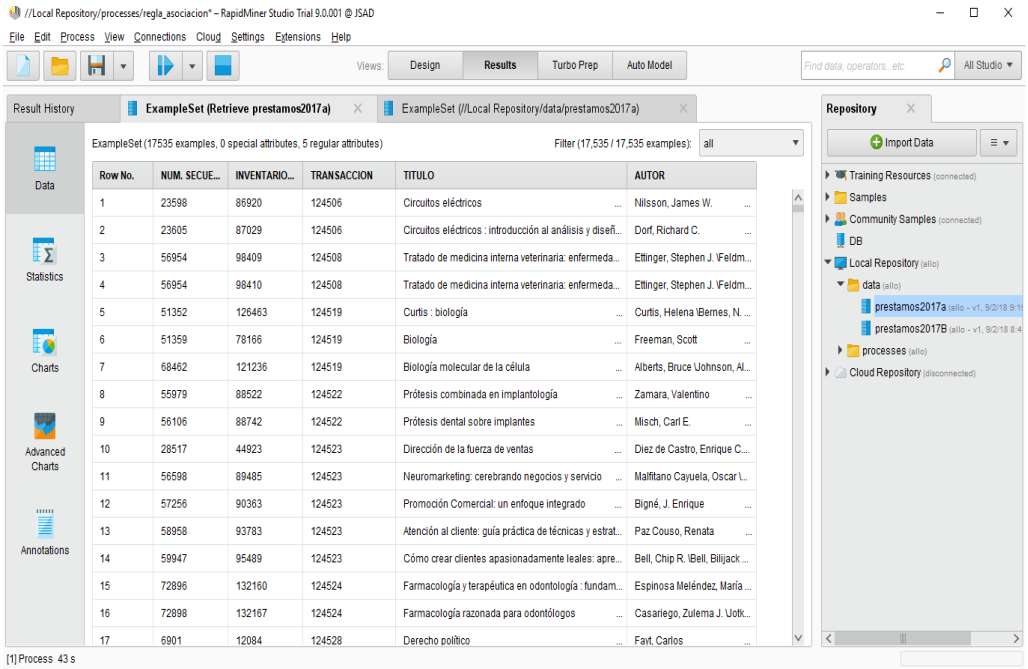


Figura 5.1 : Pre-visualización de las variables

A continuación, debemos utilizamos el operador ***"Select Attributes"***, el mismo que nos permite seleccionar, o filtrar, las columnas que deseamos utilizar (preparación de los datos). Eligiendo "subset" en la opción "attribute filter type" (ver figura 5.2), nos permitirá seleccionar los atributos, haciendo luego click en el botón "Select Attributes", que vamos a considerar para la creación de nuestro modelo. Seleccionamos los campos autor, título y transacción.

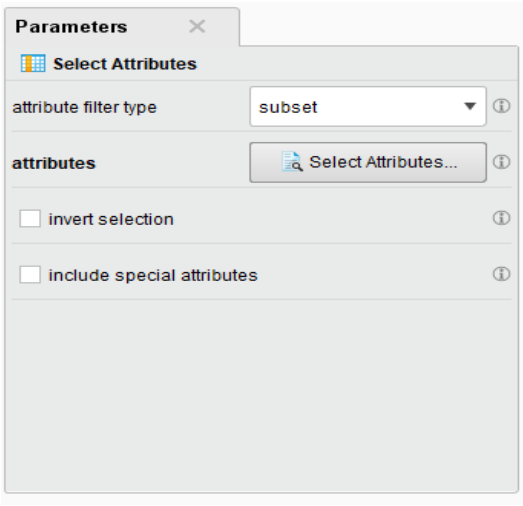


Figura 5.2 : Configuración del operador Select Attributes

Una vez ejecutado el operador, vemos que ha filtrado los campos requeridos, presentando los siguientes resultados, luego de ejecutar el operador, en su salida (ver figura 5.3).

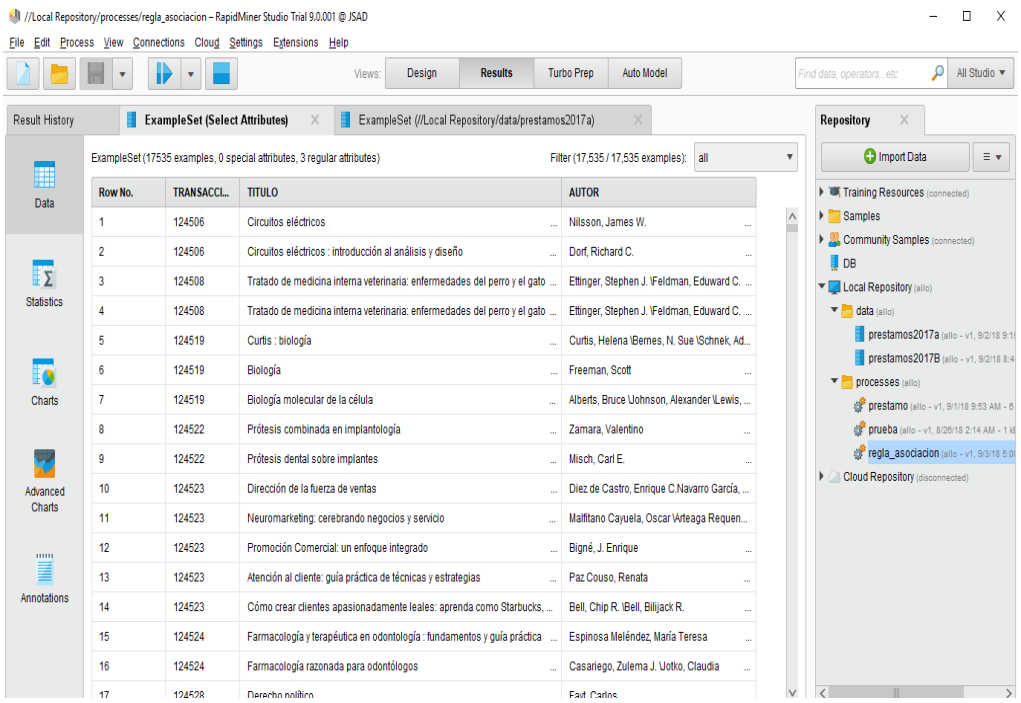


Figura 5.3 **:** Resultados del operador Select Attributes

Ahora debemos de seleccionar el operador que permitirá contabilizar los datos y realizar ciertas operaciones tradicionales de un manejador de BD SQL como agrupamientos (Group By) o funciones condicionales para grupos (Having), estas operaciones se realizan mediante el operador ***"Aggregate"***, permitiéndonos contabilizar los títulos, agrupándolos por transacción (ver figura 5.4), antes de proceder a pivotear los datos, es decir, pasarlos de filas a columnas.

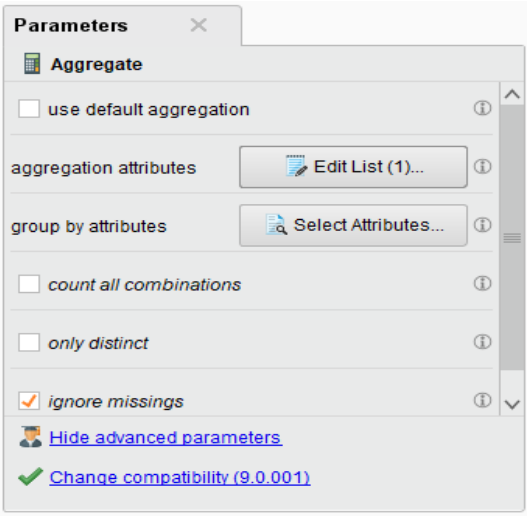


Figura 5.4 : Configuración del operador Aggregate

Luego, necesitamos mostrar los datos en forma de columnas, el operador ***"Pivot"***, nos permitirá presentar los datos (títulos) en forma de columnas, mostrándonos los ítems prestados en una determinada transacción.

A este operador, debemos indicar los parámetros por los cuales vamos a agrupar nuestras variables; para este caso agruparemos por transacción mediante la opción “group attribute” e indicaremos que el índice del atributo recae en el título del ítem, mediante la opción “index attribute” (ver figura 5.5).

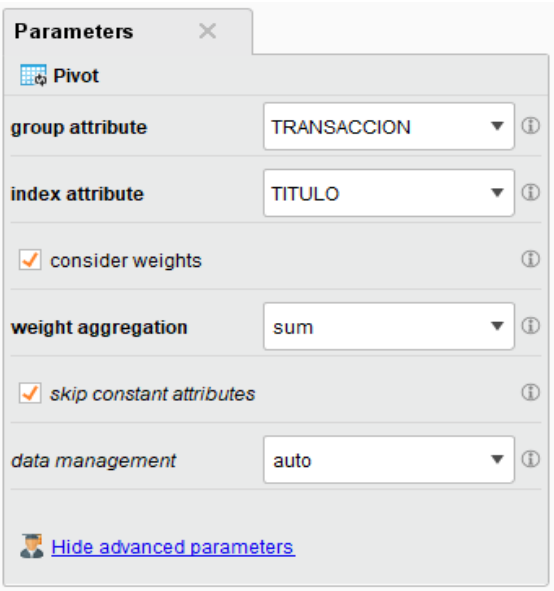


Figura 5.5 : Configuración del operador Pivot

Podemos ver cómo han rotado los títulos (ítems) agrupados por transacción, transformándolos en columnas individuales. Así también, podemos ver que aquellos títulos que no se encuentran presentes en una transacción, el operador les ha asignado un signo de interrogación (ver figura 5.6).

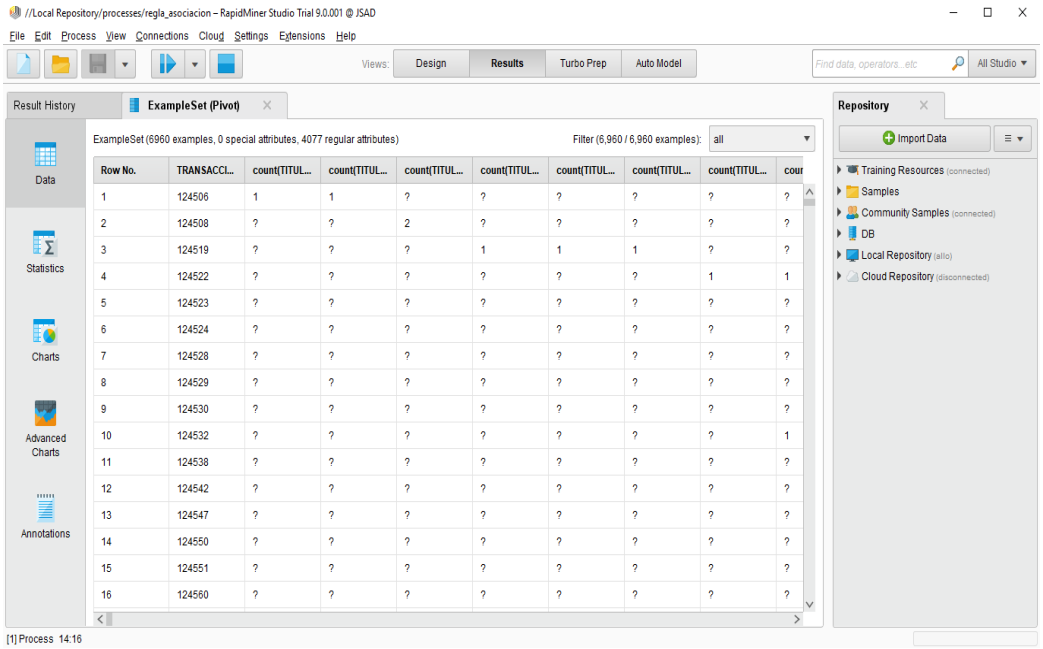


Figura 5.6 : Resultados de operador Pivot

Luego, utilizaremos los operadores ***"Rename by Replacing"*** (ver figura 5.7) para el cambio de los nombres temporales, asignados mediante la instrucción count del operador ***"Aggregate"***, por los nombres correspondientes, asignados a los atributos mediante los cuales fueron agrupados.

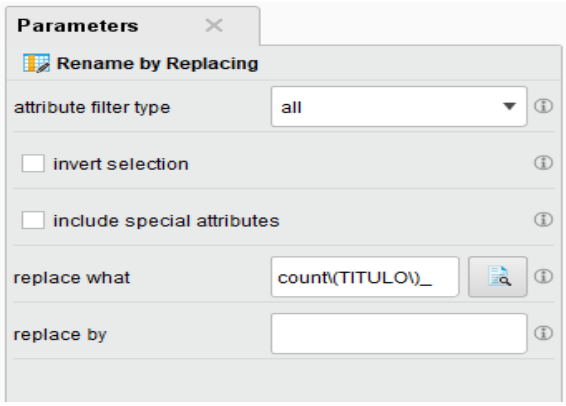


Figura 5.7 : Configuración del operador Rename by Replacing

Por su lado, el operador ***"Replace Missing Values"*** (ver figura 5.8), nos permitirá realizar la asignación de ceros a los ítems que no están presentes en determinadas transacciones.

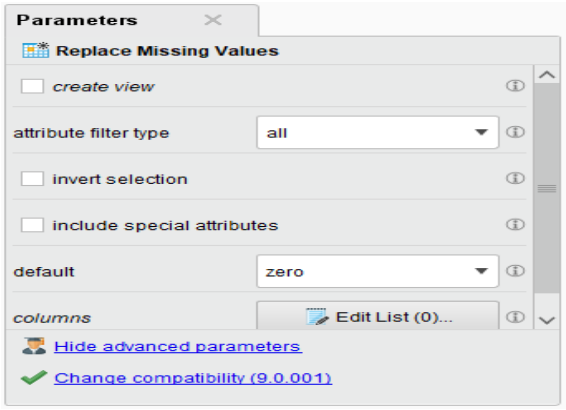


Figura 5.8 : Configuración del operador Replace Missing Values

Los resultados que obtendremos (ver figura 5.9), luego de ejecutar el operador, nos mostrará la matriz mediante ceros y unos.

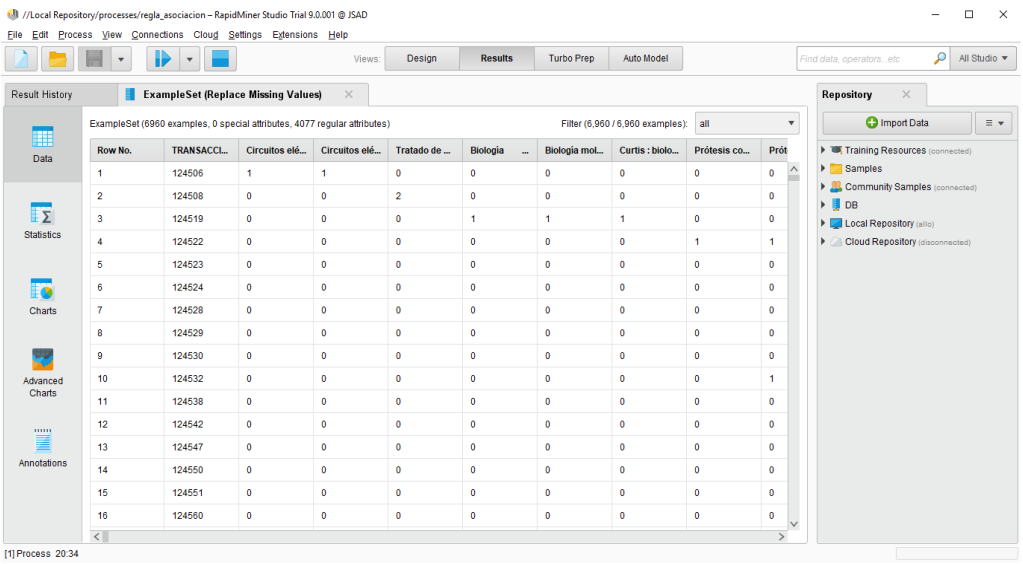


Figura 5.9 : Resultados del operador Replace Missing Values

Debido a que el modelo, no acepta ceros y unos como insumos para encontrar reglas de asociación (las reglas de asociación no realizan las operaciones mediante números), debemos realizar la transformación de variables a datos binomiales, transformar los unos en “true” y los ceros en “false.

Para lograr lo anteriormente expuesto, utilizaremos el operador ***“Numerical to Binomial”***, permitiéndonos transformar los valores numéricos a binomiales (ver figura 5.10). Al momento de realizar la selección de atributos, hemos elegido el campo transacción, y luego, hemos seleccionado la opción de “*invert selection*”, esto permitirá aplicar el operador a los diferentes títulos, sin tomar en consideración la transacción.

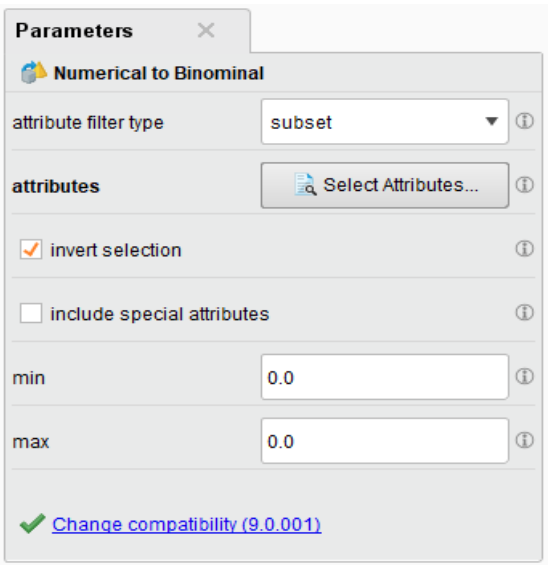
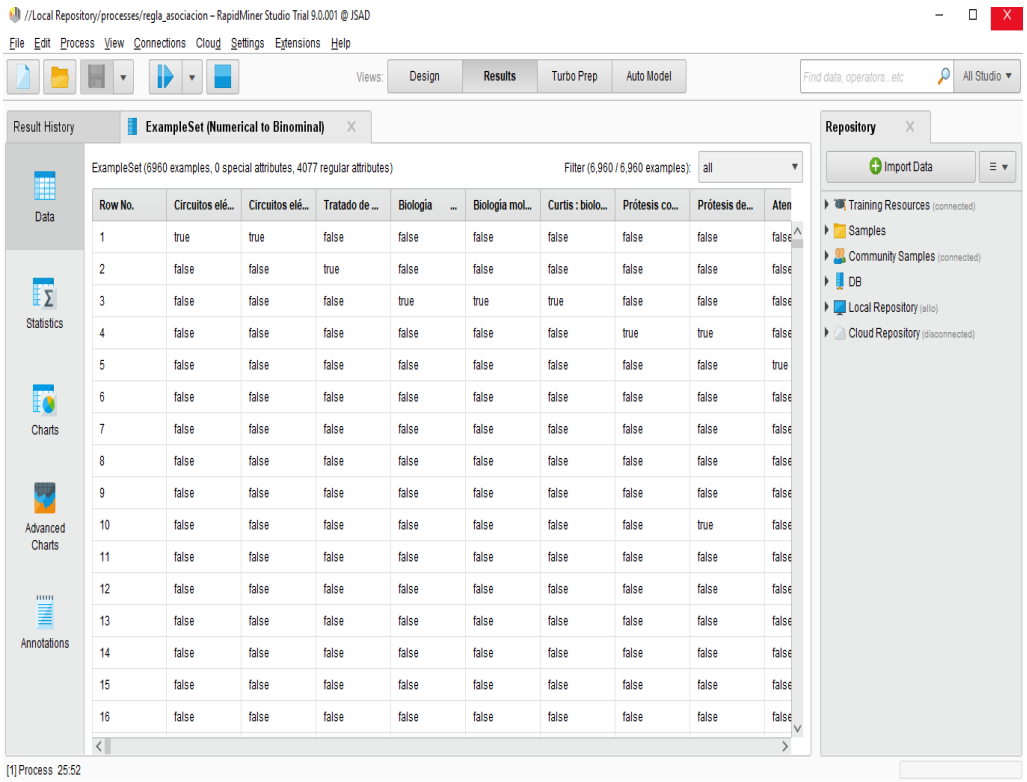


Figura 5.10 : Configuración del operador Numerical to Binomial

Como podemos apreciar, mediante la figura 5.11, se ha realizado los cambios de ceros por false (en títulos no presentes en las transacciones), y los unos por true (para títulos presentes en una transacción).



Row No.	Circuitos elé...	Circuitos elé...	Tratado de ...	Biología ...	Biología mol...	Curtis : biolo...	Prótesis co...	Prótesis de...	Aten
1	true	true	false	false	false	false	false	false	false
2	false	false	true	false	false	false	false	false	false
3	false	false	false	true	true	true	false	false	false
4	false	false	false	false	false	false	true	true	false
5	false	false	false	false	false	false	false	false	true
6	false	false	false	false	false	false	false	false	false
7	false	false	false	false	false	false	false	false	false
8	false	false	false	false	false	false	false	false	false
9	false	false	false	false	false	false	false	false	false
10	false	false	false	false	false	false	false	true	false
11	false	false	false	false	false	false	false	false	false
12	false	false	false	false	false	false	false	false	false
13	false	false	false	false	false	false	false	false	false
14	false	false	false	false	false	false	false	false	false
15	false	false	false	false	false	false	false	false	false
16	false	false	false	false	false	false	false	false	false

Figura 5.11 : Resultados del operador Numerical to Binomial

Ahora, tenemos que identificar la columna tipo “ID”, para realizar esto, utilizaremos el operador ***“Set Role”***, eligiendo la columna que servirá de identificador, en este caso, corresponde a la columna “transacción”, y lo haremos mediante la opción “attibute name” (ver figura 5.12), quedando de esta forma definida la columna principal que va a servir para el análisis de los datos.

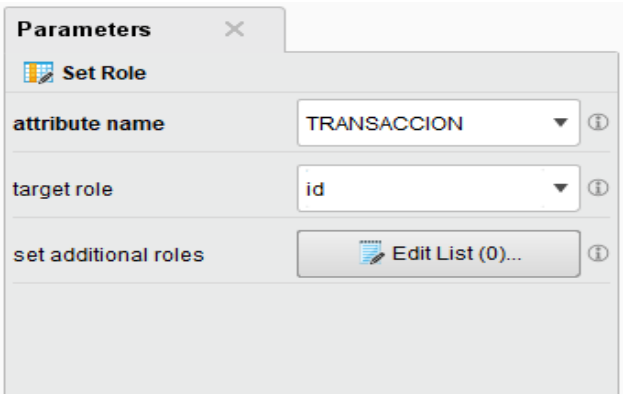


Figura 5.12 : Configuración del operador Set Role

Ejecutado el operador, podemos observar (ver figura 5.13) que la columna transacción ha sido identificada como ID, como ya lo habíamos indicado con anterioridad, es la columna principal, mediante la cual sabremos en qué transacciones están presentes un determinado número de ítems (títulos).

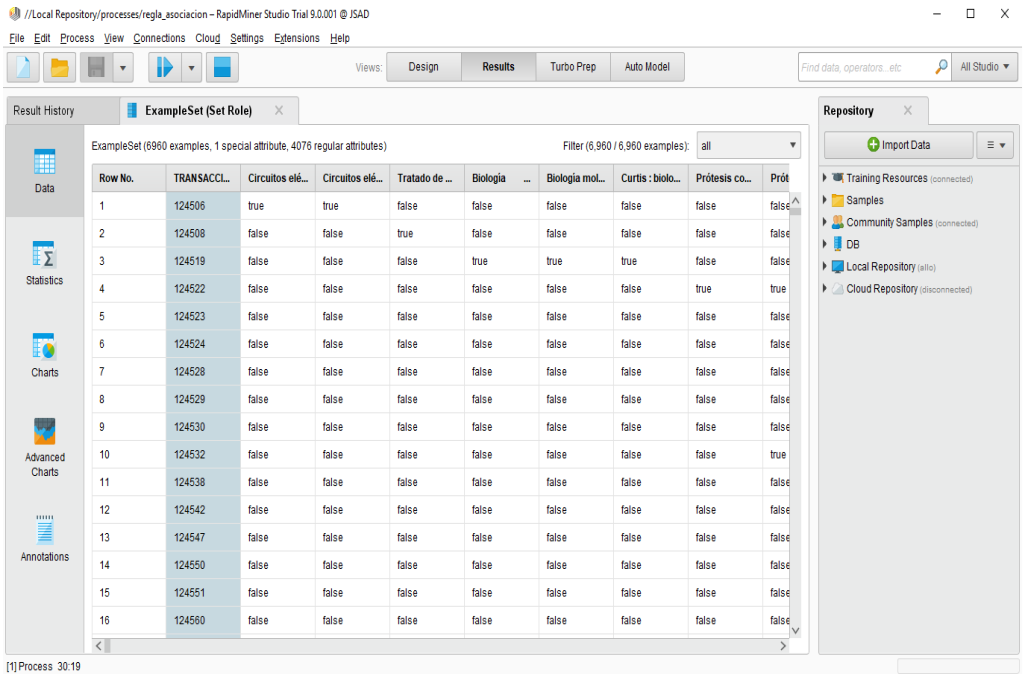


Figura 5.3 : Resultado del operador Set Role

5.2 Técnicas de visualización gráfica

Las técnicas de visualización gráficas tienen dos objetivos:

- Aprovechar la gran capacidad humana, y de los sistemas, para extraer patrones a partir de imágenes.
- Ayudar al usuario a comprensión de los patrones descubiertos por la herramienta utilizada en el proceso del DM.

A su vez, estos objetivos derivan en diferentes usos que se pueden realizar de la visualización de los datos, tales como:

- Visualización *previa*, utilizada para un mejor entendimiento de los datos y sugerir probables patrones, o incluso, la herramienta que se debe utilizar.
- Visualización *posterior*, utilizada para mostrar patrones y tener una mejor comprensión de los mismos [29].

En este proyecto se utilizará la visualización *posterior*, debido a que se pretende validar y mostrar el resultado de los datos procesados ante los expertos interesados.

Mediante la opción “Circle” (ver figura 5.14) podemos apreciar de forma gráfica la salida que presenta las diferentes reglas de asociación.

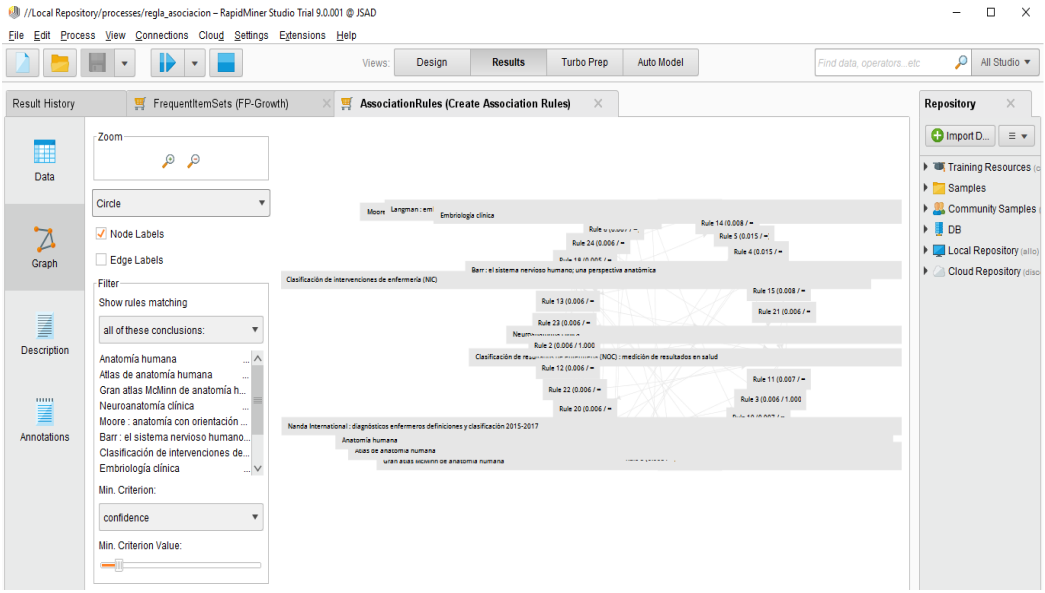


Figura 5.14 : Presentación gráfica de resultados tipo circle

5.3 Análisis de los residuos

En la implementación de este algoritmo, no existe análisis los residuos, debido a que estos se consideran dentro del modelo de regresión lineal, el cual para contrastar los resultados aplica hipótesis de linealidad, normalidad, homoscedasticidad, no autocorrelación e independencia, modelo descartado para la implementación del algoritmo en este trabajo.

Este análisis implica el uso de gráficas como el histograma de frecuencias, analizando si se ajusta, o no, a una distribución normal, revisando valores atípicos e intentando conseguir la normalidad de los residuos, mediante la representación de puntos de pares (X y Y),

detectando tendencias crecientes o decrecientes, verificando la existencia de autocorrelación o correlación serial.

5.4 Evaluación de modelos

Las diferentes técnicas de minería de datos, crea modelos de carácter descriptivo y predictivos, los predictivos responden preguntas e interrogantes, tales como, ¿Cuáles serán las ventas para el próximo año?, en cambio los modelos descriptivos permiten identificar relaciones entre los datos, basados en sus características, por ejemplo, podemos identificar si los “Clientes que compran un producto A, suelen comprar también un producto B”.

Un ejemplo frecuentemente utilizado para el último caso es el consistente en el análisis *market basket case*, el cual se utiliza frecuentemente en supermercados para determinar ítems que se compran en conjunto, algo que permitirá luego establecer su ubicación dentro de las estanterías con la finalidad de aumentar las ventas.

Para nuestro proyecto, el cual se encuentra basado en la creación de reglas de asociación, modelo clasificado dentro de los modelos descriptivos, utilizaremos el algoritmo FP-Growth (el mismo que es una versión mejorada del algoritmo Apriori), dentro del cual debemos poner especial atención a los valores de soporte y confianza, temas abordados en la sección correspondiente a la *validación del modelo*.

En las primeras pruebas realizadas, con los parámetros predefinidos por el software, con soporte de 0.95 (ver figura 5.15), no hay resultados que ayuden a la construcción del modelo (ver figura 5.16).

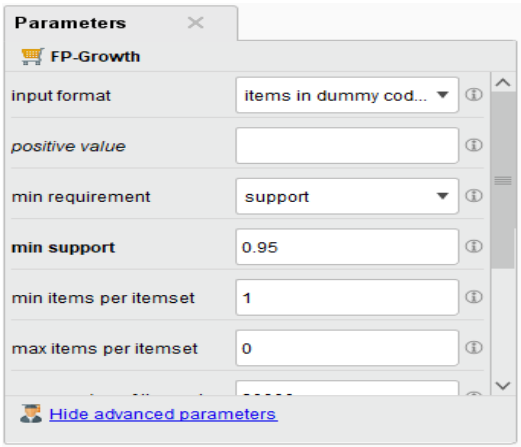


Figura 5.4 : Ejecución con parámetros Predefinidos para creación de algoritmo

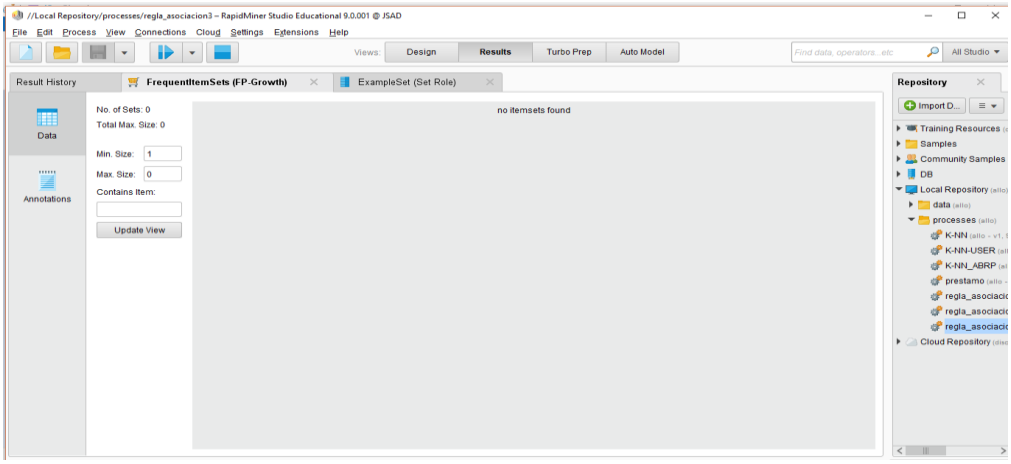


Figura 5.5 : Resultados obtenidos con parámetros predefinidos

Así también podemos observar que el operador “Create Association Rules”, con confianza mínima, predefinida de 0.8 (ver figura 5.17), tampoco muestra los resultados esperados (no ha creado reglas de asociación), tal como se muestra en la figura 5.18.

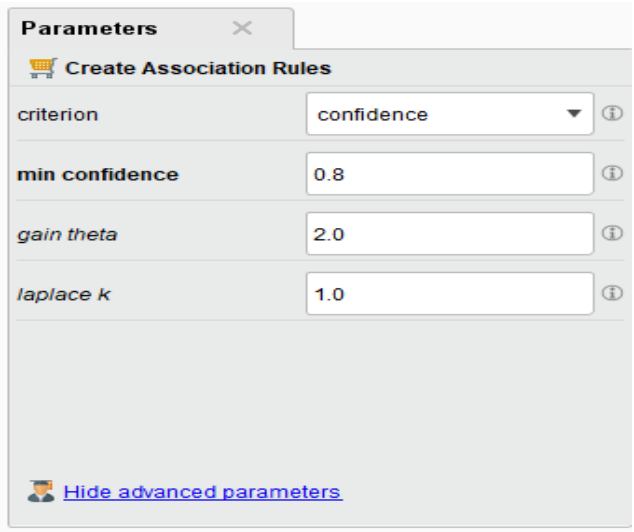


Figura 5.6 : Ejecución con parámetros predefinidos

Al igual que, en la creación del modelo, el operador para la creación de reglas de asociación no tiene los resultados esperados, por lo tanto, debemos realizar pruebas con asignación de otros valores que nos permiten obtener mejores resultados.

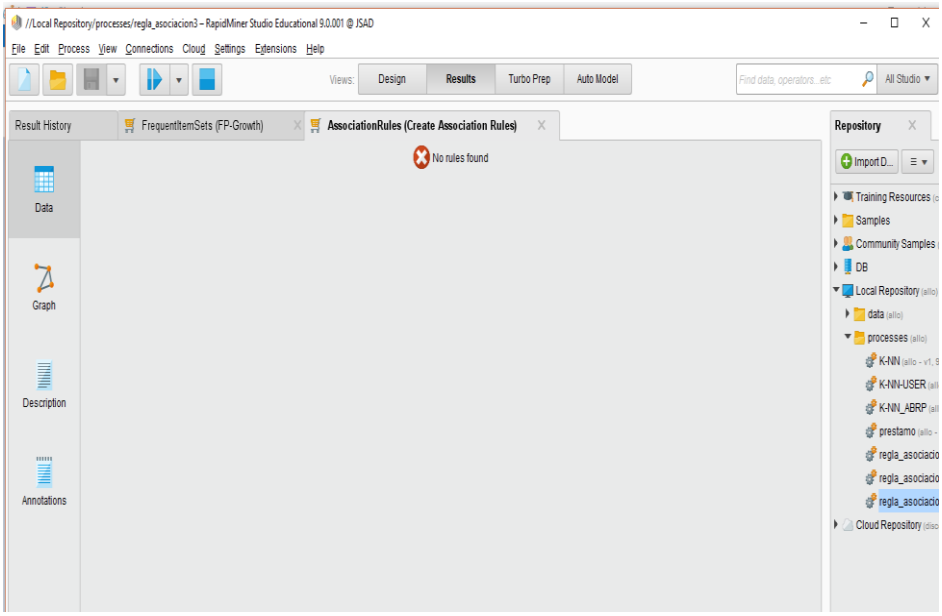


Figura 5.78 : Resultados de la ejecución con parámetros predefinidos

Debido a que, no existen resultados mediante los parámetros predefinidos por el programa, evaluaremos otros parámetros que permitan la obtención de resultados satisfactorio para el objetivo del proyecto, esto se abordará con más detalles en la siguiente sección referente a la construcción del modelo.

5.5 Construcción del modelo

Para la construcción del modelo se requiere que los datos preparados sean utilizados de forma iterativa, que se puedan aplicar algoritmos y técnicas sobre diferentes vistas “minables” con la finalidad de descubrir patrones de comportamientos sobre la utilización del material bibliográfico.

Luego de la evaluación realizada para la selección del algoritmo que permite la creación de las reglas de asociación (**FP-Growth**), seguiremos realizando nuevas pruebas que permitan la obtención de un modelo satisfactorio para responder a las necesidades de información requeridas.

Para la creación de un modelo satisfactorio, debemos probar con otros soportes mínimos (**min support**), además, debemos considerar que las transacciones con más de un título son pocos frecuentes (conjuntos de

artículos que ocurren juntos), eso se evidenció en la primera corrida con el valor de soporte predefinido en 0.95, el cual no produjo los resultados esperados.

Ahora, probaremos con un soporte de 0.005, es decir, que por lo menos en el 0.05% de las transacciones los ítems (dos o más), estén presentes en una transacción.

Así también, debemos definir los ítems mínimos y máximos por itemset, que para el caso hemos elegido 2 y 5 respectivamente (ver figura 5.19).

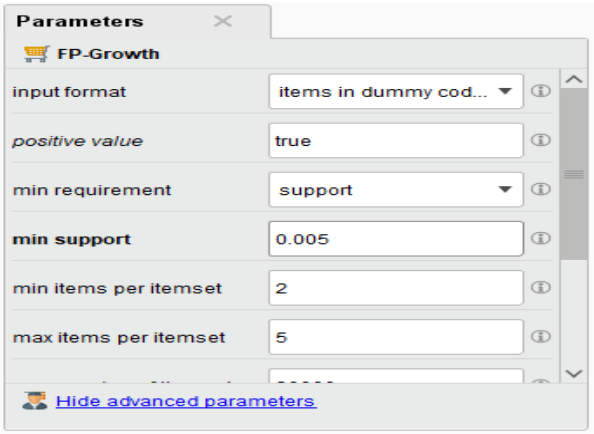


Figura 5.8 : Prueba con soporte 0.005

Observemos que el operador, que procesa el algoritmo FP-Growth, tiene dos salidas, la primera sólo nos mostrará los datos de la base transformada y la segunda es la que conformará los itemset frecuentes para la creación de las reglas de asociación (ver figura 5.20).

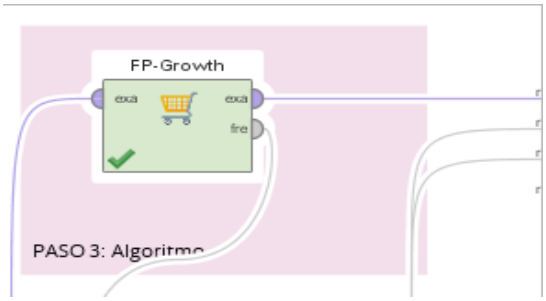


Figura 5.20 : Salidas del operador para el algoritmo FP-Growth

Con los datos proporcionados, y generados por el modelo, se crearán las diferentes "reglas de asociación", este paso lo realizaremos mediante el operador "**Create Association Rules**", en el que definiremos el criterio con el valor mínimo con el cual evaluaremos la confianza; empezaremos asignado una confianza mínima (*min confidence*) de 0.02 (ver figura 5.21), el valor predefinido (0.8) en el operador no mostrada resultados debido a que el modelo en sí no había creado los itemsets, necesarios para la creación posterior de las reglas de asociación.

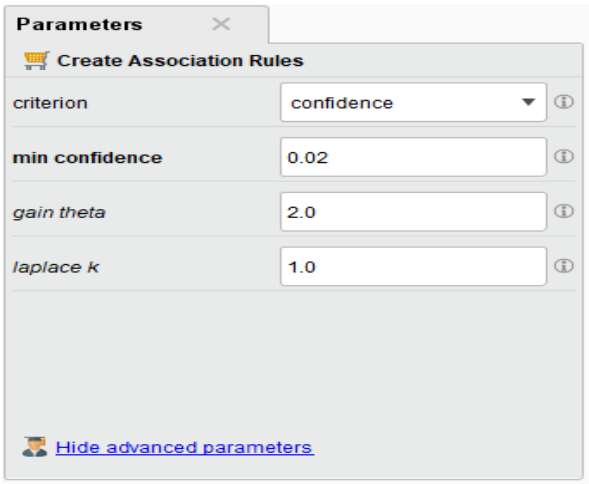


Figura 5.9 : Prueba con confianza 0.02

5.6 Implementación del modelo

Se ha desarrollado e implementado el algoritmo seleccionado, sobre la herramienta RapidMiner, la cual ha permitido realizar diferentes pruebas de desempeño para el respectivo análisis de los datos, aplicando tareas para el proceso de minería de datos.

El proceso para la obtención de las reglas de asociación, se describen mediante la figura 5.22, la cual está compuesta por las siguientes cuatro fases o pasos:

1. Pre-procesamiento, carga, unificación de datos, y selección de atributos a partir de los campos cargados.
2. Proceso de ETL, estableciendo un conjunto de operadores que permiten la extracción, normalización y transformación de los datos, así como la identificación de variables que permitan la obtención del modelo.
3. Implementación y ejecución del algoritmo, en el cual se definirán y probarán parámetros que permitan la salida de resultados que se consideren apropiados para la generación de reglas.
4. Creación de las reglas de asociación, resultados que se analizarán, se evaluaran, y se confrontaran entre los diferentes resultados obtenidos a partir de las pruebas establecida.

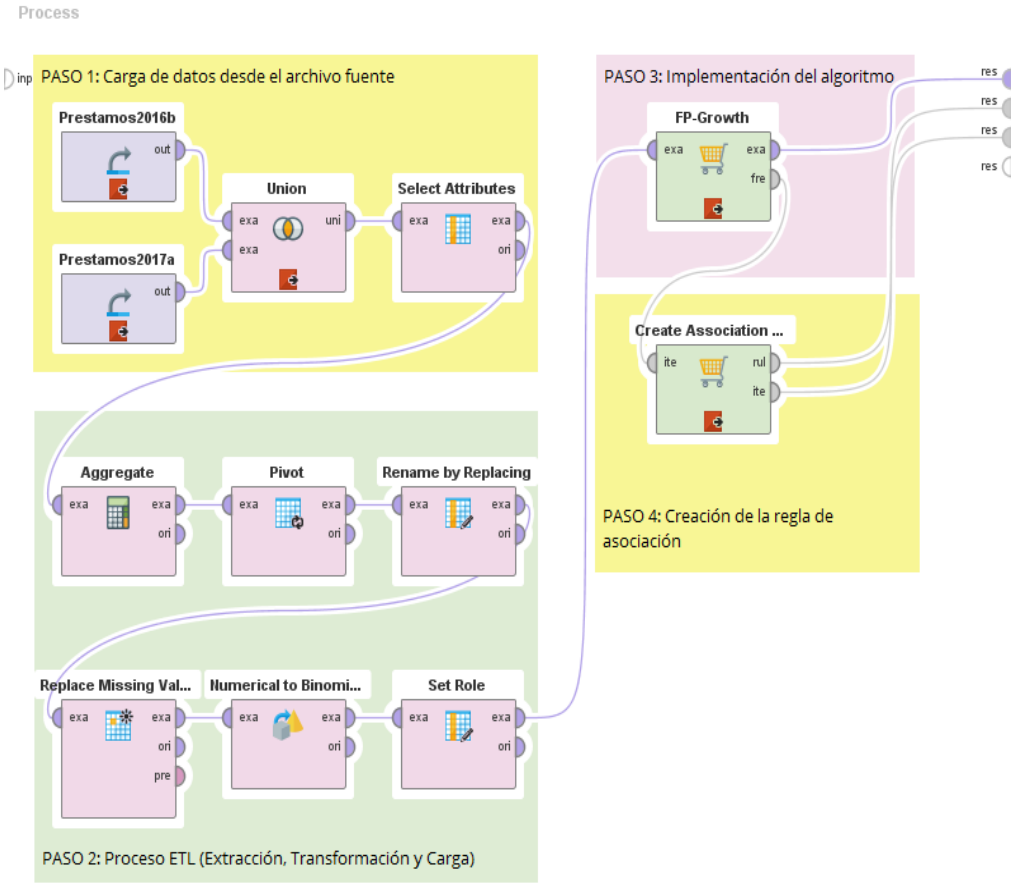


Figura 5.10 : Proceso para la implementación del modelo

Ciertos operadores contienen un punto de quiebre (breakpoint) luego de ser ejecutado, esto sirve para realizar un seguimiento detallado a los resultados proporcionados luego de la ejecución, de cada operador; en total el tiempo de ejecución ha tomado aproximadamente 2 minutos con 25 segundos, sobre un total de 33.859 registros que corresponden a las transacciones de los dos años analizados (2016 y 2017).

También debemos mencionar que, existen una serie de criterios que se deberían considerar al momento de aplicar las medidas de interés, entre las principales consideramos:

El número de atributos, las reglas con pocos antecedentes son por lo general más fáciles de comprender, en contraparte con las de muchos antecedentes que deben su aparición, posiblemente, al ruido que generan los datos o representar un caso especial.

Coste de la mala clasificación, estará vinculado a los datos con los que se trabaje, un falso positivo puede ser menos influyente que un falso negativo.

Clases de la distribución, los resultados dependerán de la división o clasificación de los datos.

Orden de los atributos, ciertos atributos serán mejores al momento de realizar discriminación de clases, es así que, si se encuentran dichos atributos en una regla, esta mejorará su posición ante las demás.

Asimetría de la medida, es mejor una medida que pueda distinguir entre antecedente y consecuente [30].

CAPÍTULO 6

ANÁLISIS DE RESULTADOS

6.1 Pruebas de la aplicación

Las pruebas realizadas, no sólo permitirán la visualización de las diferentes Reglas de Asociación, sino que mejorará la presentación y comprensión de los resultados obtenidos mediante la herramienta utilizada, en comparación con herramientas que no poseen este tipo de resultados, sino más bien, que su presentación es más textual, requiriendo conocimientos técnicos avanzados para su presentación.

Hemos procedido a realizar las pruebas, registrando diferentes parámetros para una mejor obtención de los resultados, tal como se muestra a continuación:

Tabla 24
Ejecución de la prueba uno

Prueba No. 1	
Parámetros	**Resultados obtenidos**
Soporte mínimo: 0.005 Items mínimos por itemsets: 2 Items máximos por itemsets: 5 Confianza: 0.08	Número de ítemsets: 8 Número máximo de ítems por itemsets: 2 Reglas de asociación: 0

Nota: la 1era prueba, no brindó los resultados esperados, se ejecutó con los parámetros predefinidos por el programa (ver figuras 5.16 y 5.18)

Tabla 25

Ejecución de la prueba dos

Prueba No. 2	
Parámetros	**Resultados obtenidos**
Soporte mínimo: 0.005 Items mínimos por itemsets: 2 Items máximos por itemsets: 5 Confianza: 0.08	Número de ítemsets: 8 Número máximo de ítems por itemsets: 2 Reglas de asociación: 0

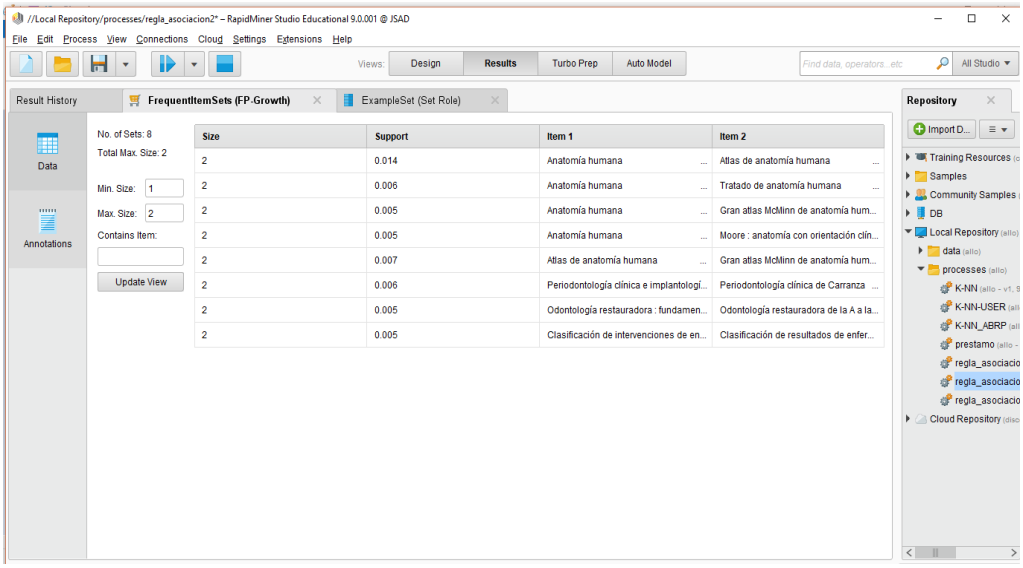


Figura 6.1 : Modelo generado en la prueba dos

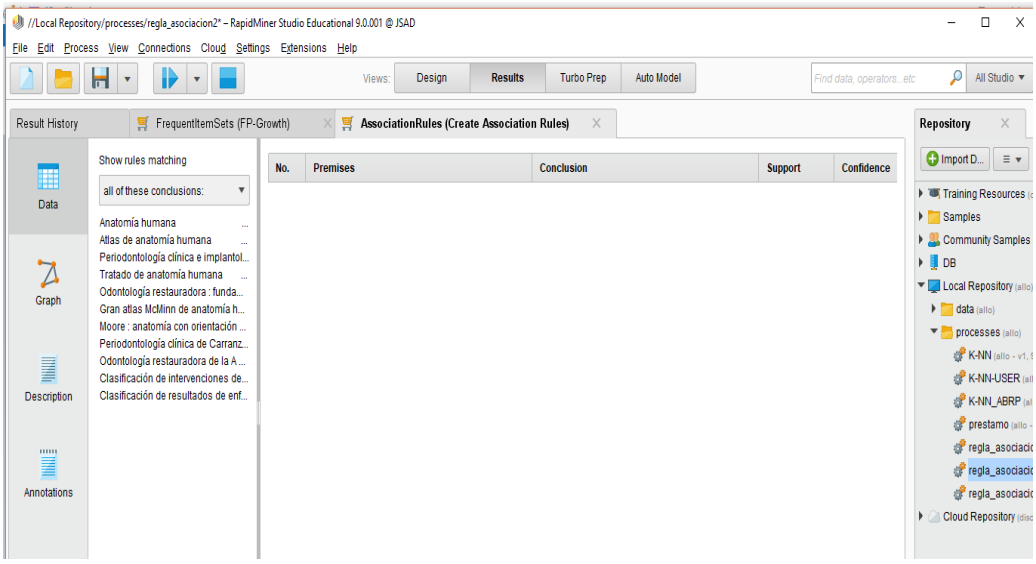


Figura 6.2 : Reglas generadas en la prueba dos

Tabla 26

Ejecución de la prueba tres

Prueba No. 3	
Parámetros	**Resultados obtenidos**
Soporte mínimo: 0.001 Items mínimos por itemsets: 2 Items máximos por itemsets: 5 Confianza: 0.01	Número de itemsets: 79 Número máximo de ítems por itemsets: 3 Reglas de asociación: 165 Soporte máximo: 0.014 (regla 7) Soporte mínimo: 0.001 (regla 2) Confianza máxima: varias reglas contienen confianza máxima Confianza mínima: 0.455 (regla 2)

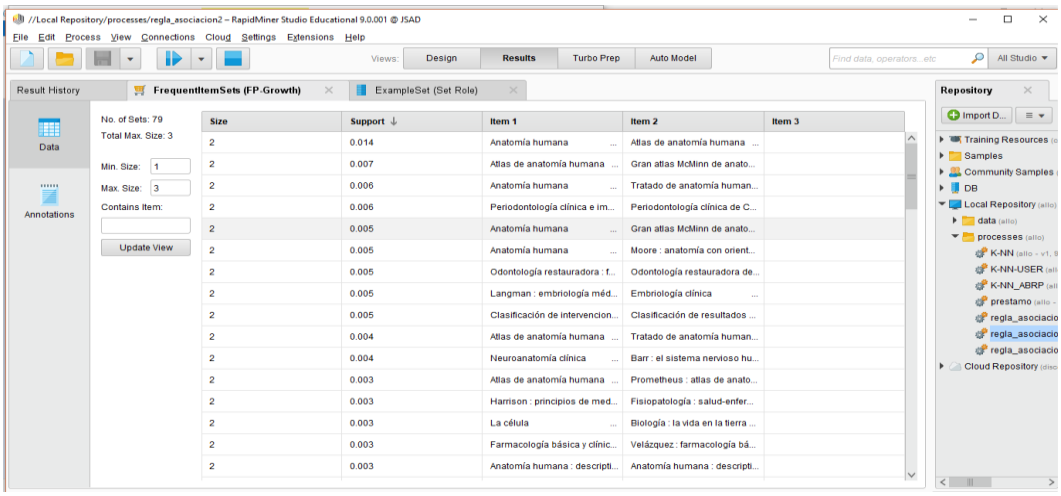


Figura 6.3 : Modelo generado en la prueba tres

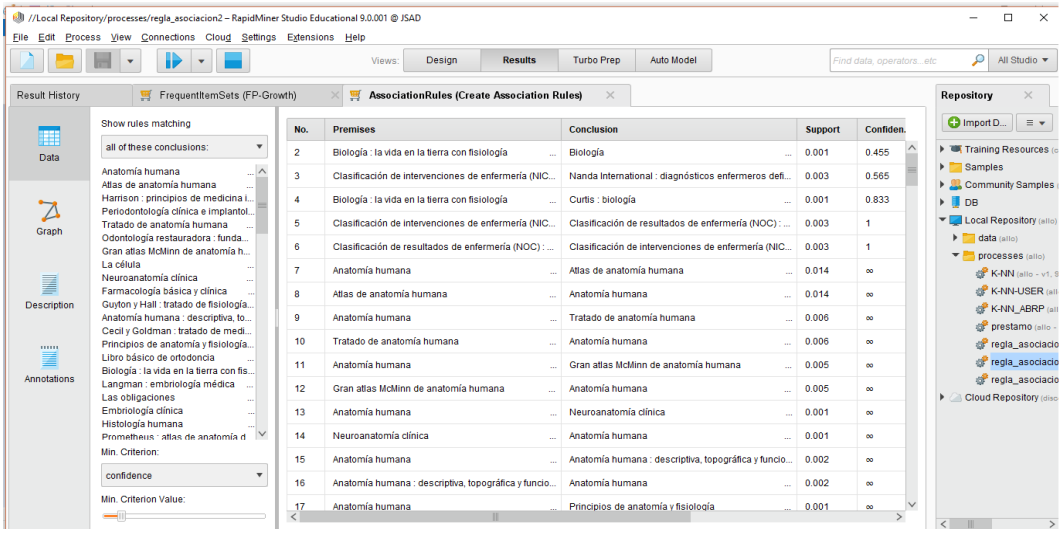


Figura 6.4 : Reglas generadas en la prueba tres

La prueba número tres proporciona mayores y mejores resultados, estos se analizarán con más detalles en el siguiente apartado.

6.2 Verificación y visualización de resultados

Debido a que la prueba número uno no proporciona resultados positivos, analizaremos los resultados proporcionados por la prueba número dos.

Podemos apreciar, por ejemplo, que cuando un usuario presta el libro “Clasificación de intervenciones de enfermería (NIC)”, está presente también el libro “Clasificación de resultado de enfermería (NOC)”, con un soporte de 0.006 y una confianza de 1. Algo parecido, cuando se realiza una transacción y solicitan el libro de “Anatomía humana”, entonces también solicitan el libro de “Atlas de anatomía humana” son un soporte de 0.015 y un nivel de confianza de 1 (ver figura 6.5).

Ba

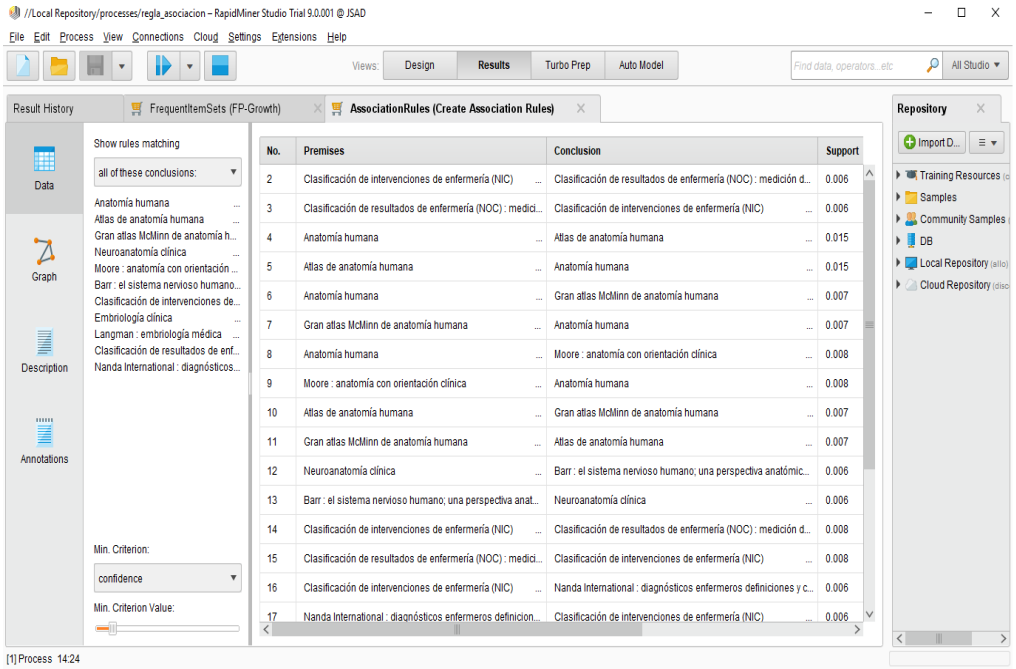


Figura 6.5 : Reglas de asociación obtenidas

Estos resultados permiten conocer la relación que existe entre ciertos temas de libros, lo que permitirá tomar algunas consideraciones al momento del desarrollo de las colecciones (sobre todo al momento de las adquisiciones).

En la figura 6.6 podemos apreciar, de forma gráfica, la relación que existen, en determinadas transacciones, entre los libros relacionados con temas como: Atlas de anatomía, Neuroanatomía clínica, Anatomía humana, Tratado de anatomía humana, entre otros, los cuales se encuentran estrechamente relacionados mediante los préstamos generados por los usuarios.

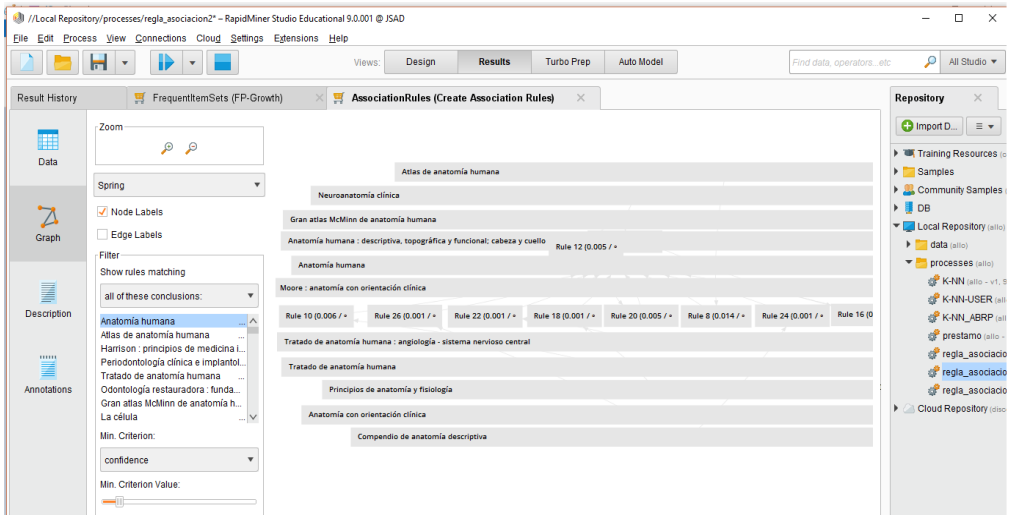


Figura 6.6 : Representación gráfica de asociación de temas de anatomía

Otro ejemplo, que podemos analizar, son los libros correspondientes al área de salud, específicamente el área de enfermería, tales como: Diagnostico enfermeros definiciones clasificación, clasificación de resultados de enfermería (NOC), y el que corresponde a Clasificación de intervenciones de enfermería (NIC), vinculados mediante las reglas 6, 122, 165,124 y 166, tal como se muestra en la figura 6.7.

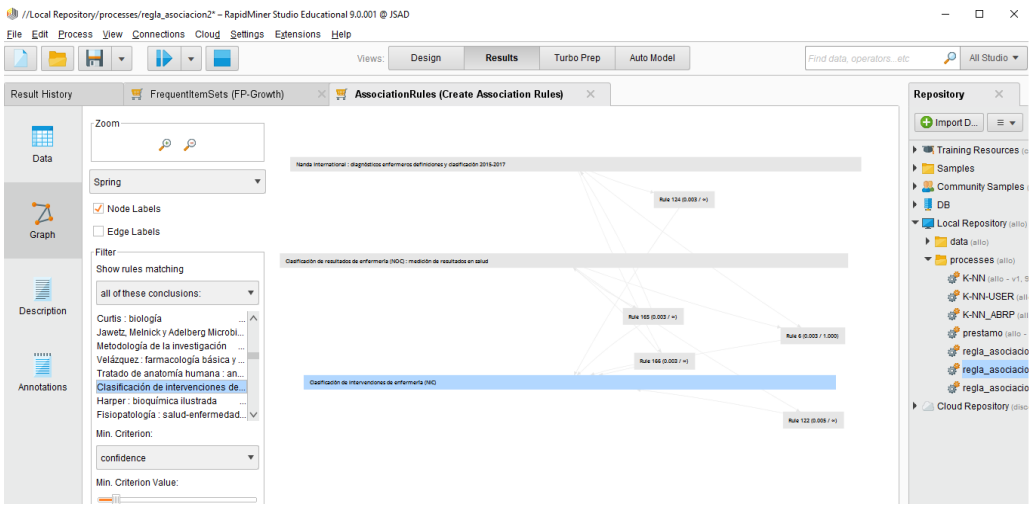


Figura 6.7 : Representación gráfica de asociación de temas de enfermería

6.2.1 Beneficios de la implementación de la MD para la institución

Dentro de los principales beneficios que obtendrá la institución, al momento de aplicar técnicas de minería de datos, podemos mencionar las siguientes:

- **Ahorro a largo plazo**, el proceso la implementación no es difícil ni costoso, sobre todo para realizar pruebas exploratorias de datos, permitiendo en breve obtener información y generar conocimiento confiable.
- **Toma de decisiones estratégicas**, basadas en el conocimiento generado, permitirá acelerar la toma de decisiones, por ejemplo, realizar compras de material bibliográfico de manera más precisa, pues se conoce la rotación de los ítems, y la relación que existe entre ellos al momento de realizar una transacción.
- **Relación con el cliente**, el tiempo de respuesta para la prestación de servicios es menor, además, se puede ofrecer, como valor agregado, ítems que el usuario desconoce, pero que sirven de ayuda para sus estudios (ítems que han sido utilizado por otros usuarios con similares características).
- **Cumplimiento con normativa**, permitirá cumplir con indicadores nacionales e internacionales, referentes a los

procesos relacionados con las normativas reguladoras referente a la atención a usuarios (indicadores de gestión de servicios bibliotecarios).

6.3 Difusión, uso y monitorización

El proceso de creación del modelo es una parte fundamental del proyecto, pero para que tenga la efectividad que se necesita, se debe difundir la información generada entre los interesados, expuesta o expresada de forma sencilla y comprensible, pues este conocimiento pasa a ser parte del know-how de la institución; la información es uno de los bienes intangibles más preciados por las instituciones.

Su uso dependerá de las condiciones y necesidades operacionales, por ejemplo, dentro de los resultados identificados se tienen una serie de ítems de similares características que son motivos de préstamos en conjunto, por lo tanto, se debería evaluar la posibilidad de ubicar en estanterías cercanas los libros relacionados con el área correspondiente.

Si bien es cierto, el almacenamiento en las bibliotecas está basado en un modelo de clasificación decimal DEWEY[3], se debe considerar, ubicar

[3] Sistema numérico decimal creado para la organización de los libros en las bibliotecas, creado por Melvil Dewey en 1876, divide el conocimiento en diez grandes categorías, que a su vez, se pueden dividir en varias subcategorías.

de forma más cercana aquellos ejemplares que poseen una mayor rotación, así como, aquellos que están presente en transacciones de más de un ejemplar.

El modelo creado, no puede ser estático, debe revisarse de forma periódica, a fin de identificar posibles cambios en las variables; cambios que pueden estar validados por diferentes factores como el económico, requerimientos de organismos de control, la competencia, fuentes de datos, entre otros.

Una monitorización oportuna, permitirá al modelo ser revalidado de forma frecuente a fin de identificar posibles desviaciones en su comportamiento, por lo que, puede ser necesario una actualización de su diseño.

Es aconsejable desarrollar un plan que permita la supervisión y mantenimiento de los datos a fin de alimentar y actualizar el modelo desarrollado, pudiéndose establecer procesos como:

- Selección, extracción y almacenamiento de los datos de forma semestral, en formato texto.
- Establecer procesos que permitan respetar el anonimato de los usuarios, protegiendo su integridad e información sensible.

- Los archivos deben ser almacenados de forma cronológica, de acuerdo con el semestre analizado.
- Cada periodo modelado y analizado, debe poseer distintas gráficas que permitan una rápida interpretación de los resultados. El uso de infografías ayudaría con este objetivo.
- Deben crearse informes semestrales y anuales con la finalidad de ser entregados a los diferentes interesados.

Finalmente, se debería elaborar una normativa que permita la colaboración de los diferentes involucrados, a fin de que se faciliten los datos necesarios, además, de ayudar a que los informes se elaboren y difundan de forma independiente, y no necesariamente por el requerimiento de alguien en particular, haciendo partícipe a toda la comunidad académica.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

1. La minería de datos proporciona información oculta, que permite un mejor análisis de la información, ayudando a mejorar la toma de decisiones.
2. La aplicación del modelo de reglas de asociación, mediante pruebas de ensayo y error, nos ayuda a identificar patrones asociativos relacionados con las transacciones.
3. Los modelos, así como los parámetros con los que se evalúan, no pueden ser estáticos.

4. Los datos almacenados no producen conocimiento, esto se logra cuando se procesan de una forma adecuada y oportuna.
5. Existen mecanismos y software que permiten el procesamiento de datos, independientemente de origen y del formato en que se guarden.
6. Los servicios pueden mejorar al aplicar técnicas adecuadas de minería de datos.

Recomendaciones

1. Experimentar el uso de otras técnicas de minería de datos que contribuyan a otras necesidades de información.
2. Implantar un prototipo basado en técnicas de minería de datos dentro de los Sistemas Bibliotecarios con la finalidad de mejorar la experiencia del usuario.
3. Revisar o establecer políticas y procedimientos que ayuden a la recolección, procesamiento de los datos, así como, la correcta y oportuna difusión de los resultados.
4. Los resultados deben segmentarse y ponerse a consideración de los interesados, representándose de una forma clara y precisa.

BIBLIOGRAFÍA

[1] Hernández Orallo J, Ramírez Quintana MJ, Ferri Ramírez C. Introducción a la minería de datos. Madrid: Pearson; 2010.

[2] Hasperué W. Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas. Editorial de la Universidad Nacional de La Plata (EDULP); 2014.

[3] Haehnel AR. Utilización de base de datos temporales en sistemas de prestaciones médicas. Universidad de Palermo, 2014.

[4] Amazon. ¿Qué es una base de datos documental? – AWS. Amaz Web Serv Inc 2018. https://aws.amazon.com/es/nosql/document/ (accessed July 5, 2018).

[5] Silberschatz A, Korth HF, Sudarshan S. Fundamentos de bases de datos (5a. ed.). Madrid, SPAIN: McGraw-Hill España; 2006.

[6] Reyes F, C S, Ruiz Lobaina M. Minería Web: un recurso insoslayable para el profesional de la información. ACIMED 2007;16:0–0.

[7] Espino Timón C. Análisis predictivo : Técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso 2017. http://openaccess.uoc.edu/webapps/o2/handle/10609/59565 (accessed February 22, 2018).

[8] Hernández Orallo J. Curso Almacenes de Datos y Minería de Datos n.d. http://users.dsic.upv.es/~jorallo/cursoDWDM/index.html (accessed February 22, 2018).

[9] Escarcega B. Introducción a la Minería de Datos. Bus Intell Data Wareh Monterrey México Gravitar 2007. https://gravitar.biz/bi/data-mining-intro/ (accessed February 22, 2018).

[10] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Amsterdam ; Boston : San Francisco, CA: Elsevier ; Morgan Kaufmann; 2006.

[11] López R. La minería de datos, entre la estadística y la inteligencia artificial. Min Datos Expr Inf 2015. https://comunidad.iebschool.com/bigdata/2015/05/13/la-mineria-de-datos-entre-la-estadistica-y-la-inteligencia-artificial/ (accessed February 28, 2018).

[12] Universidad de Costa Rica. ¿Cuál es la relación entre Big Data, minería de datos y Estadística? Univ Costa Rica n.d. https://www.ucr.ac.cr/noticias/2016/02/25/cual-es-la-relacion-entre-big-data-mineria-de-datos-y-estadistica.html (accessed May 21, 2018).

[13] Romero P. El uso fraudulento de datos de Facebook pone en peligro investigaciones académicas 2018. http://www.publico.es/ciencias/facebook-fraudulento-datos-facebook-pone-peligro-investigaciones-academicas.html (accessed May 9, 2018).

[14] ESOMAR. Código Internacional ICC/ESOMAR para la práctica de la Investigación de Mercados, Opinión y Social y del Análisis de Datos 2016.

[15] UNESCO. Código de Ética para la Sociedad de la Información, propuesto por el Consejo Intergubernamental del programa Información para Todos (PIPT), Paris: 20111.

[16] Franganillo J. Implicaciones éticas de la minería de datos. Anu ThinkEPI 2010. http://eprints.rclis.org/14881/ (accessed May 8, 2018).

[17] Código Orgánico de la Economía Social de los Conocimientos. 2016.

[18] Riquelme JC, Ruiz R, Gilbert K. Minería de Datos: Conceptos y Tendencias. Intel Artif Rev Iberoam Intel Artif 2006;10.

[19] Ferruccio MA, García Alonso AI, Gómez SX. Minería de Datos, Sao Paulo: 2004.

[20] Justicia de la Torre MC. Nuevas técnicas de Minería de Texto: Aplicaciones. Granada: 2017.

[21] Gutierrez JE. Descubrimiento de conocimientos en la base de datos académica de la Universidad Autónoma de Manizales aplicando Redes Neuronales. Universidad Atónoma de Manizales, 2012.

[22] Logicalis. Cómo elegir sistema de minería de datos 2014. https://blog.es.logicalis.com/analytics/como-elegir-sistema-de-mineria-de-datos (accessed May 10, 2018).

[23] Rodríguez Suarez Y, Díaz Amador A. Herramientas de Minería de Datos 2009. http://www.redalyc.org/html/3783/378343637009/ (accessed May 10, 2018).

[24] 5 de los mejores software de minería de datos de Código Libre y Abierto | El rincón de JMACOE n.d. http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/ (accessed May 14, 2018).

[25] Piatetsky G, KDnuggets. KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll: RapidMiner Continues To Lead 2014. https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html, https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html (accessed August 24, 2018).

[26] Trujillo JC. Diseño y explotación de almacenes de datos: conceptos básicos de modelado multidimensional. Alicante, SPAIN: ECU; 2013.

[27] Pinho Lucas J. Métodos de clasificación basados en asociación aplicados a sistemas de recomendación 2010.

[28] Riveros M, Alejandra M, Elida Beguerí G. Reglas de Asociación con los datos de una biblioteca universitaria. Rev Cuba Cienc Informáticas 2015;9:30–45.

[29] Orallo JH, Ramírez Quintana MJ, Ferri C. Curso de Doctorado “Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del

Software” 2007. http://users.dsic.upv.es/~jorallo/docent/doctorat/ (accessed September 20, 2018).

[30] Ruiz Jiménez MD. Modelado formal para representación y evaluación de reglas de asociación. Tesis Doctoral. Universidad de Granada, 2018.