ESTADO DEL ARTE DE LA MINERÍA DE DATOS APLICADA A LA INTELIGENCIA DE NEGOCIOS

DUBER MAURICIO CONTRERAS PABON

TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE MAGISTER EN GESTIÓN DE PROYECTOS INFORMÁTICOS

DIRECTOR:

PhD: JOSÉ ORLANDO MALDONADO BAUTISTA

UNIVERSIDAD DE PAMPLONA

FACULTAD DE INGENIERAS Y ARQUITECTURA

MAESTRÍA EN GESTIÓN DE PROYECTOS INFORMÁTICOS

PAMPLONA

2018

Dedicatoria

Dedicado a mi familia, mis padres, por su gran apoyo y el ánimo a seguir siempre adelante, mi esposa y hermano, por siempre estar ahí.

Agradecimientos

Agradezco a Dios, por darme nuevas oportunidades de vivir, salud y fortaleza para poder llegar a esta etapa y culminar satisfactoriamente.

Resumen

Este documento presenta una completa revisión bibliográfica sobre trabajos desarrollados, las últimas tendencias, técnicas y aplicaciones de la minería de datos enfocada a la inteligencia de negocios; una revisión teórica que estableció el conocimiento generalmente aceptado, se efectuó exploración de la literatura generada en los últimos 5 años, tanto a nivel nacional como internacional. La revisión fue organizada y analizada desde diferentes puntos de vista, como el tiempo (orden cronológico), las fuentes de datos, selección, exploración y visualización de los datos. Una categorización de los aportes en relación con los modelos predictivos y descriptivos en torno a la minería de datos, categorizando técnicas orientadas a: clasificación, clustering, regresión y reglas de asociación. Fue posible establecer el estado actual del área de estudio y así mismo plantear posibles trabajos futuros, potenciales campos de aplicación, oportunidades de negocio y líneas de profundización e investigación.

Palabras claves: minería de datos, inteligencia de negocios, técnicas de minería de datos, herramientas de minería de datos.

Abstract

This document presents a complete bibliographical review on developed works, the latest trends, techniques and applications of data mining focused on business intelligence; a theoretical review that established the generally accepted knowledge, was carried out exploration of the literature generated in the last 5 years, both nationally and internationally. The review was organized and analyzed from different points of view, such as time (chronological order), data sources, selection, exploration and visualization of the data. A categorization of the contributions in relation to the predictive and descriptive models around data mining, categorizing techniques oriented to: classification, clustering, regression and association rules. It was possible to establish the current status of the study area and also propose possible future work, potential fields of application, business opportunities and lines of research and deepening.

Keywords: data mining, business intelligence, techniques of datamining, tools of datamining.

Tabla de contenido

Lista de figuras

Listado de siglas

(BPM): procesos de gestión de negocios.

(GEP): programación de expresión genética.

(ACO): Optimización a través de colonias hormigas.

(CCSDMS): sistema de minería de datos de correlación de coeficiente de ventas.

(SCE): software de estimación de costo.

(COCOMO): modelo de construcción o estimación de costo

(RTBI): inteligencia de negocios en tiempo real.

(DSS in cloud): sistema de soporte de decisiones en la nube.

(RMS): estrategias de marketing relacional.

(LCRM): modelo de regresión de clases latentes.

(MPM): mercadeo móvil personalizado.

(CRM): gestión de relación con el cliente.

(BMs): métodos de negocios.

(OM): minería de opinión.

(SA): análisis de sentimientos.

(KBS): sistemas basados en conocimiento.

(SLR): revisión sistemática de literatura.

(DSR): investigación en ciencias del diseño.

(ASD): desarrollo ágil de software.

(FCMs): mapas cognoscitivos difusos.

(RPDS): sistema de dispensación de prescripción.

(PHC): comunicaciones de salud pública.

(SMEs): pequeñas y medianas empresas.

(MIS): sistema de gestión de información.

(MkIS): sistemas de información de marketing.

(ASM): minería de reglas de asociación.

(PRM): modelo de regresión probabilística.

(L-HMM): modelo de Markov oculto lexicalizado.

(CRF): modelo de campos aleatorios condicionales.

(FARM): minería de reglas de asociación difusa.

(LSA): método de análisis semántico latente.

MACOM: multi-agente basado en mapas cognitivos difusos.

(KEA): métodos de extracción de frases.

(ACE): método de extracción de conceptos claves.

(ICE): extractor de conceptos mejorado.

(H-MK-SVM): modificación de máquina de soporte vectorial kernel múltiple jerárquica mejorada.

(MLPNN): red neural de percepción multicapa.

K-NN-IR: K vecino más cercano mejorado para tratamiento de reglas de inducción.

K-means-IR: K-means modificado para el tratamiento de reglas de inducción.

(MIA): minería de agentes inteligentes.

(PAM): minería de asociación principal.

(RApriori-TdMl): algoritmo para el descubrimiento de reglas de asociación multinivel.

(SKU): código de identificación de productos.

(MOD. DIRICHLETH): modelo de distribución a priori, basado en estadística bayesiana.

(MET. CHAID): Detección de interacción automático Chi cuadrado.

(ALG. FITMOS): Minería objetos frecuentes.

(WEKA): Entorno para análisis de conocimiento Waikato.

(HER. KEEL): Extracción de conocimiento basado en aprendizaje evolutivo.

(DATAWAREHOUSE): almacén de datos.

(GIRVAN-NEWMAN): algoritmo de búsqueda de comunidades, elimina los nodos con la mayor centralidad de inserción, identifica comunidades de interés.

Lista de tablas

Introducción

En la medida en que la tecnología y el mundo han venido evolucionando también lo han hecho muchos sistemas que permiten al hombre realizar tareas de clasificación, predicción, organización y estructuración de la información, de manera tal que se pueda extraer y valorar aspectos relevantes de la misma. La minería de datos ha surgido como una alternativa y/o estrategia muy apreciada en cuanto a la organización, clasificación y aprovechamiento de la información en referencia a la escala de descubrimiento del conocimiento, el almacenamiento y extracción de información de grandes volúmenes de datos.

En áreas específicas como la Inteligencia de Negocios ha sido un espacio en donde las técnicas usadas por la minería de datos aportan en gran manera al desarrollo de la misma, mediante la aplicación de metodologías, técnicas y herramientas en donde permite que esta rama recopile, extraiga y analice de manera eficiente la información y así poder tener a disposición aspectos claves para el mercadeo, las ventas, sistemas de análisis de crédito, clasificación de perfiles de los consumidores, tipos de compradores, comportamiento de pago entre muchos otros aspectos más de relevancia en cuanto al sector se refiere, lo que constituyen la inteligencia de negocios.

En el presente trabajo se realizó un recorrido minucioso por la bibliografía disponible en las bases de datos de conocimiento científicamente aceptado y otras fuentes de información con el objetivo de precisar cuáles han sido las técnicas, los métodos, los modelos, las metodologías, los autores, los logros y dificultades que se presentaron dentro del proceso de aplicación y evolución de la minería de datos en la inteligencia de negocios, se organizó el sumario a partir del análisis realizado y se estableció una prospectiva de lo que pudiera ser la aplicación de estas herramientas, se establecen nuevas líneas de desarrollo e investigación tomando como base los aportes más relevantes en los últimos 5 años del uso de este tipo de técnicas y herramientas de minería de datos.

El presente documento está estructurado en tres capítulos, en el primer capítulo se encuentran los aspectos preliminares del documento; en el segundo capítulo, se desarrolla el marco teórico y conceptual en donde se establecen los referentes académicos e investigaciones que comprenden el fundamento de los objetos de estudio; en el tercer capítulo, se establece el desarrollo del estado del arte de la minería de datos aplicada a inteligencia de negocios, finalmente se plantearon conclusiones, recomendaciones y trabajos futuros producto de la clasificación y análisis de la información.

Capítulo 1. Preliminares

1.1 Presentación del problema y Justificación

La pregunta sobre la cual se planteó el desarrollo del presente proyecto se originó en precisar que investigaciones, avances, aplicaciones y novedades se han desarrollado por parte de la comunidad científica, académicas y particulares respecto a la minería de datos como herramienta para la toma de decisiones, apoyo, clasificación y elección de determinados elementos en el área de inteligencia de negocios.

Es fundamental para el fortalecimiento de esta y otras áreas de estudio así como para el aporte de nuevo conocimiento que se realicen recopilaciones documentales en las cuales se plasme los avances, descubrimientos, requerimientos, dificultades, nuevos procedimientos y futuras tendencias. Efectuando un breve recorrido por la literatura asociada se puede vislumbrar que existe una carencia relacionada con trabajos o escritos que recopilen y documenten los aportes que se han realizado en pro del avance de las mismas disciplinas, de ahí la relevancia de una investigación en donde se recopile los aportes más importantes y a su vez se documenten y clasifiquen los mismos, de esta manera posibilite definir el estado de estas áreas y los posibles enfoques de trabajos futuros.

La inteligencia artificial, la inteligencia de negocios y la minería de datos, son campos del conocimiento que avanzan rápidamente. Para formular problemas y proyectos de investigación que aporten avances significativos en el área y se consoliden como aportes de nuevo conocimiento, se hace necesario revisar de manera exhaustiva los últimos aportes encontrados en la literatura, con lo cual se puede determinar la frontera del conocimiento en el área. Este es el punto de partida para poder hacer aportes que pueden ser valiosos para la comunidad científica nacional e internacional. Por otra parte, la revisión permitió explorar soluciones planteadas a problemas en el área de inteligencia de negocios, replicables en un entorno nacional y local, que pueden ser implementados como soluciones tecnológicas que generan ideas nuevas para el emprendimiento empresarial.

1.2. Objetivos

A continuación se describirán el objetivo general y los específicos que formaron parte del desarrollo del proyecto.

1.2.1. Objetivo general.

Precisar que investigaciones, avances, aplicaciones y novedades se han desarrollado por parte de la comunidad científica, académicas y particular respecto a la minería de datos como herramienta para la toma de decisiones, apoyo, clasificación y elección de determinados elementos en el sector del comercio, negocios y ventas.

1.2.2. Objetivos específicos.

- Revisar el marco teórico y conceptual en los campos del conocimiento de la Inteligencia de negocios y la minería de datos.
- Recopilar y documentar los reportes, artículos y ponencias encontrados en diferentes repositorios científicos y académicos para su posterior análisis.
- Analizar la documentación recopilada para su posterior organización desde diferentes perspectivas, y su presentación como aporte original del proyecto.
- Clasificar la documentación recopilada y analizada en categorías surgidas del proceso de consulta y análisis.

1.3. Metodología para realizar el estado del arte

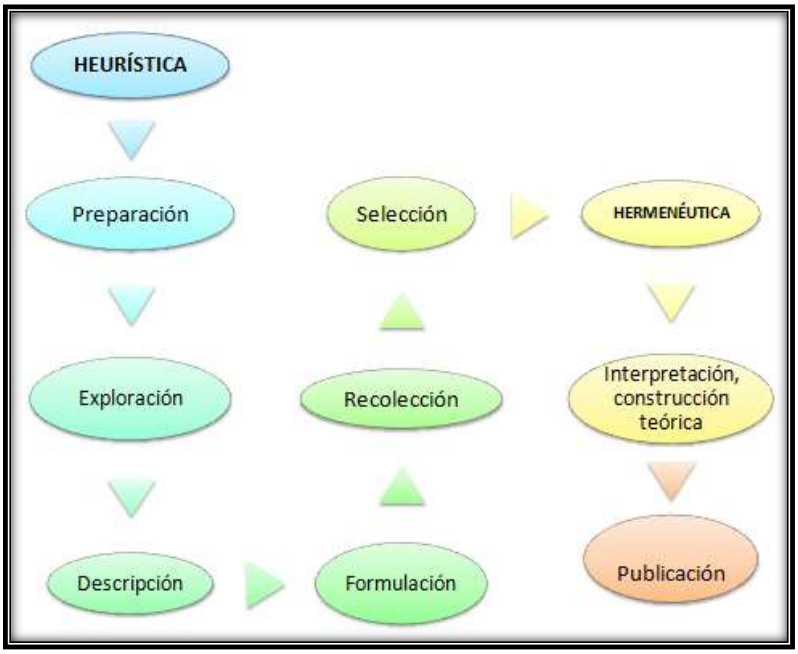
El estado del arte se puede definir como una modalidad de la investigación documental que permite el estudio del conocimiento acumulado escrito dentro de un área específica; su finalidad es dar cuenta del sentido del material documental sometido a análisis, con el fin de revisar de manera detallada y cuidadosa los documentos que tratan sobre un tema específico. Esto significa que es una recopilación crítica de diversos tipos de texto de un área o disciplina, que de manera escrita, formaliza el proceso cognitivo de una investigación a través de la lectura de la bibliografía hallada durante la indagación del problema así como los temas y los contextos. (Londoño et al., 2014).

Es de gran relevancia que quien desee desarrollar un estado del arte se plantee algunos interrogantes como es el caso de los campos de indagación, la consulta de información que se ha definido mantengan una relación directa con el tema de la investigación. Además de lo anterior es importante identificar conceptos fundamentales y aportes dentro de los documentos que han sido seleccionados, con base a esto construir el estado del arte, así priorizar en estas áreas o dimensiones. (Londoño et al., 2014).

La construcción de un estado del arte debe estar bien fundamentada y basada en criterios que permitan y posibiliten cumplir el objetivo del desarrollo de mismo, según lo describe (Londoño et al., 2014), Hoyos (2000) están basados en los fines que se persiguen (finalidad), en hallar una estructura que le dé unidad (coherencia), en el respeto y la ética del investigador frente al manejo de los datos (fidelidad), en lograr una unidad dentro de la diversidad de los documentos analizados (integración), en alcanzar un resultado final en el que se demuestre una visión de totalidad de los fundamentos teóricos como conjunto (comprensión).

Siempre es necesario tener una guía o una ruta mediante la cual se facilite la construcción del estado del arte y para esto (Londoño et al., 2014), define algunas fases para la elaboración de un estado del arte, viendo desde el punto de vista de la hermenéutica y la heurística, la ruta de describe en la figura 1. Siendo el inicio de la misma la concepción heurística, pasando por la preparación, exploración, descripción, formulación, recolección y selección y desde la perspectiva hermenéutica la formulación, la construcción teórica y la formulación.

Figura. 1. Ruta para la elaboración del estado del arte



Fuente: (Londoño et al., 2014).

Capítulo 2. Minería de datos

2.1. Generalidades

La minería de datos es una área que puede considerarse interdisciplinar que tiene similitud con los procesos vividos cotidianamente, por ejemplo, se puede relacionar con la minería de oro o de rocas, una aproximación más cercana a la descripción del nombre realmente seria minería de conocimiento a través de los datos o quizá de una manera más corta como minería de conocimiento.

En un breve recorrido por la literatura relacionada al área de minería datos, según (Rajaraman & Ullman, 2011), plantea que la definición más comúnmente aceptada de la misma es la que hace referencia al descubrimiento de modelos de datos. Sin embargo, desde otra perspectiva, la minería de datos según (Tien, 2014), comprende aplicaciones, ingeniería y aspectos científicos. Desde la perspectiva de la Fundación Nacional de Ciencia, la describe como "amplia, diversa, compleja, longitudinal y/o establecimiento de datos distribuidos generados de instrumentos, sensores, transacciones de internet, email, video y todas otras aplicaciones y fuentes disponibles hoy y en el futuro" (NFS, 2012). Sin embargo como se evidencia no hay un consenso unificado en cuanto al concepto de la minería de datos.

La minería de datos tiene como objetivo ayudar a comprender el contenido presente en una base de datos, los cuales pueden estar almacenados en repositorios o provenir de otras fuentes, esta comprensión se puede dar de una manera genérica o de forma más específica, partiendo desde el almacenamiento y tratamiento de los datos hasta la aplicación de estos en la solución de un problema específico como la predicción de perfiles o patrones enfocados en sectores particulares. Teniendo en cuenta que los datos cuando están en las bases, son solo datos, pero cuando hay presencia de un significado o un usuario les asigna este, pasan de ser datos y se transforman en información, cuando se realiza o se predice un modelo se aplica un análisis a esta información, es entonces allí cuando se hace referencia a conocimiento. Las técnicas usadas para modelar las operaciones y descubrir conocimiento es lo que se conoce como data mining o minería de datos. (Pei, Kamber, & Jiawei, 2012).

Hoy día, la minería de datos está cobrando relevancia creciente en empresas y organizaciones para resolver problemas complejos de negocio, según (Barrientos, 2013) esta es una herramienta de alta calidad y fiabilidad para este tipo de áreas de trabajo, donde se hace necesario el manejo y análisis adecuado de bastos volúmenes de datos, con el fin de descubrir comportamientos o predecir conductas en cierto tipo de mercado o negocios.

2.2. Disciplinas con las cuales se relaciona la minería de datos

La minería de datos puede de cierta manera ser considerada de forma transversal, teniendo en cuenta que no se ciñe a una sola disciplina, sino que por el contrario esta puede ser aplicada a múltiples disciplinas. De la misma manera existen disciplinas que fundamentan y nutren la minería

de datos como es el caso de las ciencias de la computación, la estadística, el reconocimiento de patrones y el aprendizaje automático.

Según cita (Peter Denning, 2005.) “La disciplina de la computación es el estudio sistemático de procesos algorítmicos que describen y transforman información: su teoría, análisis, diseño, eficiencia, implementación y aplicación. La pregunta fundamental subyacente en toda la computación es, ¿Qué puede ser (eficientemente) automatizado?”. De esta disciplina en conjunto con la minería de datos se desprenden diversas aplicaciones, herramientas y técnicas orientadas a automatización y rendimiento de los procesos.

Teniendo en cuenta el trabajo de (ALUJA, 2001) se identifica la minería de datos y dos disciplinas muy relevantes, estas son la estadística y la inteligencia artificial específicamente mediante máquinas de aprendizaje; estas máquinas vistas como una rama de la inteligencia artificial encargada del diseño y la aplicación de algoritmos de aprendizaje. Por otra parte, la estadística, vista como una rama de la matemática aplicada a datos de observación que puede ser orientada a estudios de población, de variaciones y el estudio de los métodos de reducción de datos.

2.3. Tipos de datos y fuentes de datos

Los tipos de datos a extraer provienen de diversas fuentes, pueden ser datos estructurados, como las bases de datos, o no estructurados como la web u otros repositorios, dependiendo del dominio sobre el cual se está operando.

Cuando se hace referencia a tipos de datos estructurados se hace alusión a toda aquella información que se encuentra en la mayoría de las bases de datos, son archivos que generalmente tienen un formato o estructura definida, se suelen mostrar en filas y columnas con títulos, son datos que pueden ser ordenados y procesados de forma fácil por la mayoría de las herramientas de minería de datos. La información de datos no estructurados generalmente es la que más valor posee y está representada principalmente en forma de texto, generalmente son datos binarios que no tienen estructura interna identificable. (Pei et al., 2012).

Los datos relacionales son aquellos que contiene una serie de tablas a las que se les asigna un nombre de manera única, cada una contiene un conjunto de atributos almacenados en grandes filas, cada fila contiene la identificación de un objeto con su respectiva clave y suministra un conjunto de valores propios; los que se encuentran en almacenes de datos que básicamente son repositorios de información recopilada de múltiples fuentes, guardados bajo un esquema unificado y usualmente en un solo sitio; los datos transaccionales, pueden ir desde la manera como un consumidor compra sus artículos, hasta el número de clic que utiliza en determinada página web. (Pei et al., 2012).

Un patrón o estructura de datos puede ser definido como una serie de variables que se presentan de forma constante y que se pueden identificar dentro de un conjunto de datos más grande, un patrón es de interés si cumple cuatro condiciones: en primer lugar, que sea de fácil entendimiento,

en segundo lugar, una validación con cierto grado de certeza, en tercer lugar, que sea potencialmente útil y en cuarto lugar, que sea novedoso y original. Un patrón resulta de interés si valida una hipótesis que el usuario buscaba confirmar. (Pei et al., 2012).

Los patrones se pueden agrupar dentro de la minería en dos categorías: descriptiva y predictiva; La minería descriptiva caracteriza las propiedades de los datos en virtud del objetivo de un conjunto de datos, la minería predictiva se enfoca en una inducción de los datos actuales en función de obtener predicciones. Según cita (Espino Timón, 2017) "Los modelos descriptivos cuantifican las relaciones entre los datos de manera que es utilizada a menudo para clasificar clientes o contactos en grupos. A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular".

2.4. Tipos de modelos

Los algoritmos, técnicas y herramientas de minería de datos habitualmente crean modelos que son de dos tipos: predictivos y/o descriptivos, en el caso de un modelo predictivo se relaciona con una respuesta a preguntas sobre datos futuros. Por otra parte, un modelo de tipo descriptivo intenta proporcionar información sobre las relaciones entre los datos y sus características.

2.4.1. Modelos descriptivos.

Desde el punto de vista de los modelos de tipo descriptivo, según (Hand et al., 2001), pueden ser descritos como: "*un modelo descriptivo presenta, en forma conveniente, las características principales de los datos. Es esencialmente un resumen de los datos, que nos permite estudiar los aspectos más importantes de los datos sin ser oscurecidos por el tamaño del conjunto de datos*".

Algunas de las técnicas aplicadas en este tipo de modelos se orientan hacia la minería de patrones frecuentes, se tienen en cuenta la correlación y asociación como métodos más usados, los patrones frecuentes permiten el descubrimiento de interesantes asociaciones y correlaciones dentro de este tipo de modelos.

2.4.2. Modelos predictivos.

Los modelos de tipo predictivo son descritos según (Beltrán Martínez, 2003) de la siguiente manera: "*Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases:*

- *Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y,*
- *Prueba (prueba del modelo sobre el resto de los datos)*".

Para procesos de análisis predictivo se pueden aplicar técnicas de clasificación, regresión y reglas de asociación; en la clasificación se encuentra una función o un modelo que describe y distingue clases de conceptos, este puede ser representado de varias formas, como reglas de clasificación, árbol de decisión, fórmulas matemáticas o redes neuronales. (Hand et al., 2001). Teniendo en cuenta las definiciones anteriormente mencionadas los dos tipos de modelos interactúan entre sí con el objeto de proporcionar un mejor rendimiento de los procesos de minería.

2.5. Conceptos fundamentales de la minería de datos

Son diversas y en algunos casos muy generales las definiciones en cuanto a minería de datos se refieren, sin embargo, es posible afirmar que el núcleo de los procesos de minería está en dos grandes ramas, la predicción y la clasificación, las tareas de predicción y clasificación son también llamadas patrones de descubrimiento, estos patrones han llegado a convertirse en elementos claves para el análisis de negocios en grandes empresas y todo tipo de problemática en general. (Pei et al., 2012).

El esquema de modelamiento de los procesos de minería de datos se plantea sobre la definición de un propósito u objetivo, siguiendo con la obtención de los datos, la exploración y la limpieza de los mismos, la determinación de las tareas de minería, la selección de los métodos de minería de datos, la aplicación de los métodos y la obtención de un modelo final, hasta llegar a la evaluación del rendimiento y el despliegue o la visualización de los resultados. Los anteriores pasos conforman la secuencia de aplicación que en cuanto al proceso de minería de datos se lleva a cabo, con algunas variaciones posibles teniendo en cuenta el objetivo de aplicación de las técnicas de minería de datos. (Pei et al., 2012).

2.5.1. Pasos en el proceso de minería de datos.

En un proceso de minería de datos no es solo de importancia el adecuado entendimiento de los algoritmos y modelos utilizados sino que también es muy relevante el entendimiento del problema ya que los errores más serios se comenten comúnmente en este aspecto, este entendimiento debe ser desarrollado antes de introducirse en los aspectos propios del algoritmo que se pretenda usar. Los pasos más comunes utilizados en la minería de datos son los siguientes: (Pei et al., 2012).

- Desarrollo y entendimiento del propósito del proyecto de minería de datos: cuál es el problema, los interesados, el uso de los resultados, etc.
- La obtención de los datos usados para el análisis: involucra las diferentes fuentes de bases de datos, pueden ser internas o externas.
- Exploración, limpieza y pre procesamiento de los datos: verificación de las condiciones de los datos, rangos de valores aceptables, valores atípicos.
- Reducción de los datos: solo si es necesario.

- Determinación de las tareas de minería de datos: clasificación, predicción, clustering, etc.
- Para el caso de tareas supervisadas, bien sea para tareas de clasificación o predicción, la partición de los datos, se deben realizar tres grupos: entrenamiento, validación y prueba.
- Selección de la técnica de minería de datos a ser utilizada: regresión, redes neuronales, jerárquicas, clustering.
- Uso del algoritmo para realizar la tarea: es un proceso iterativo en donde se prueban múltiples variantes del mismo algoritmo seleccionando diferentes variables u opciones con el algoritmo.
- Interpretación de los resultados: involucra un proceso de selección del mejor algoritmo desarrollado donde es posible probar el conjunto de datos final para tener una idea del posible funcionamiento.
- Desarrollo del modelo: este paso involucra la integración del modelo en un sistema operacional ejecutando registros reales con el objeto de producir una acción o decisión.

2.5.2. Aprendizaje.

Como en todos los procesos que involucran el desarrollo de algoritmos y herramientas computacionales el primer acercamiento está relacionado con los procesos realizados por animales o seres humanos, partiendo de esta idea el aprendizaje puede ser descrito como el resultado de experiencias que se transforman en conocimiento, el ejemplo más típico de esto es el de las ratas (roedores), las cuales aprenden a identificar cebos con veneno, cuando estos animales perciben un olor o aspecto nuevo en los alimentos primero prueban cantidades muy pequeñas, posteriormente la decisión de comer del alimento es basada en la reacción que ocasiono el ingerir aquella pequeña cantidad, si el efecto es negativo el alimento es asociado con una enfermedad, de forma clara existe un mecanismo de aprendizaje que relaciona las experiencias con la ingesta de ciertos alimentos. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

Por otra parte, para hacer referencia a un proceso de aprendizaje automático se puede pensar en una máquina que aprende a distribuir correos electrónicos no deseados, una solución simple seria programar dicha máquina para que reconozca y memorice las etiquetas de correo electrónico no deseado marcadas previamente por el usuario, posteriormente cuando llegue un nuevo correo electrónico la maquina lo asociará con el conjunto de correos con características similares, así lo marcará como no deseado. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

En el momento de introducir tareas de aprendizaje a una maquina se debe estar consciente de proporcionar principios claros y bien definidos que protejan el programa de alcanzar conclusiones sin sentido o inútiles. El desarrollo de tales principios es un elemento central objetivo de la teoría del aprendizaje automático (Ben-David & Shalev-Shwartz, 2014). Dentro del aprendizaje

automático existen diversos tipos ligados con las tareas y paradigmas propios de aprendizaje, estos van relacionados con aprendizaje supervisado y no supervisado, aprendices activos y pasivos, en línea y protocolo de aprendizaje por lotes.

Un rasgo distintivo entre las técnicas de minería de datos se presenta entre los métodos de aprendizajes supervisados y no supervisados, estos son usados para tareas de clasificación y predicción. Para esto es de relevancia que se tenga a disposición los datos de interés, un ejemplo de aprendizaje supervisado puede orientarse al conocimiento de compras realizadas y no realizadas o la predicción de un tumor. Un ejemplo de un algoritmo de aprendizaje no supervisado puede ser el agrupar datos similares para luego optimizar búsquedas. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.5.3. Selección de datos.

Dentro de un proceso de selección de datos intervienen múltiples factores como: saber de dónde provienen, con qué calidad se presentan, la existencia o no de datos incompletos o que se puedan considerar como ruido así como la relación entre las variables. Las fuentes de donde provienen los datos por lo general son heterogéneas y múltiples por lo que es recomendable realizar una adecuada selección de datos. De esta manera las mediciones y análisis que se realizan a los datos son esenciales para el buen funcionamiento de los algoritmos, herramientas de minería y la obtención de buenos resultados. (Pei et al., 2012).

Existen por ejemplo técnicas de resumen de datos descriptivos, las cuales permiten identificar valores típicos de los datos y resaltar los datos con ruido o valores atípicos. Mediciones como la tendencia central y la medición de la dispersión son características esenciales cuando se desea conocer los datos, para esto existen algunas técnicas basadas en la estadística que ayudan a los procesos de selección, estas incluyen: media, mediana, modo, rango medio, mientras que las medidas de dispersión incluyen cuartiles, rango intercuartil, diagrama de cajas y varianza (Pei et al., 2012).

Existen también algunos métodos gráficos de selección de datos apoyados en la estadística como histogramas, gráficos de cuartiles, parcelas q-q, parcelas de dispersión y curvas de loess. Además existen rutinas de limpieza de los datos las cuales se encargan de manejar los datos a fin de completar datos inconclusos, suavizar el ruido y la identificación y tratamiento de los valores atípicos; en el caso de valores faltantes de los datos existen algunos métodos como: ignorar la tupla (se realiza cuando hace falta una etiqueta a una clase, especialmente en problemas de clasificación), completar el valor que falta manualmente (aunque es una técnica que requiere de mucho tiempo y no presenta un buen comportamiento en grandes volúmenes de datos), uso de una constante global para completar el valor faltante (se trata de asignar una etiqueta definida a todos los datos faltantes, aunque esto podría confundir al programa de minería teniendo en cuenta que todos estos tendrían una etiqueta de valor común, es un método simple pero no fiable), usar la media aritmética de los atributos para completar el valor faltante, usar la media del atributo para

todas las muestras que pertenezcan a la misma clase, usar el valor más probable para completar el valor faltante (puede ser determinado por regresión o herramientas basadas en inferencia). (Pei et al., 2012).

Cuando se habla de ruido en los datos se refiere a un error aleatorio en una variable medida, en cuanto al tratamiento del ruido existen algunas técnicas que permiten suavizar este ruido y dar manejo a los datos, existen técnicas de suavizado por intercalación, regresión y clustering, en ese orden, las técnicas de intercalación suavizan el valor de los datos consultando los valores a su alrededor y se distribuyen en una cantidad de depósitos, este es un método de suavizado local. En cuanto a las técnicas de regresión los datos son ajustados a una función, implica encontrar la mejor línea de ajuste para las variables de manera tal que un atributo pueda usarse para predecir otro. En el caso de valores atípicos estos pueden ser detectados mediante clustering, donde los valores similares se asignan a grupos, de esta manera los valores que quedan fuera de los grupos son considerados como valores atípicos. (Pei et al., 2012).

Otro método que ayuda a la obtención de datos de buena calidad es el análisis de relevancia, utilizado en forma de análisis de correlación y selección de subconjuntos de atributo, se puede usar para detectar atributos que no contribuyen a la clasificación o a la tarea de predicción, teniendo en cuenta que muchos de los atributos de los datos pueden ser redundantes el análisis de correlación es posible identificar si dos atributos están estadísticamente relacionados, de esta manera al encontrar dos atributos con una fuerte correlación alguno de los dos podría ser eliminado, reduciendo la redundancia. Las bases de datos también pueden contener datos irrelevantes para lo cual la selección de subconjuntos de atributos permite encontrar un conjunto reducido de atributos de modo que la distribución de probabilidad resultante sea lo más parecida a la distribución original haciendo uso de todos los atributos. (Pei et al., 2012).

2.5.3.1. Reducción de datos y reducción de dimensión.

El comportamiento y el desempeño de los algoritmos de minería de datos se pueden mejorar cuando se limita el número de variables con las cuales trabaja dicho algoritmo. El proceso de consolidación de una gran cantidad de registros o de casos en conjuntos de datos más pequeños es a lo que se le denomina reducción de datos, los métodos más generalmente usados en la reducción de datos se refieren a agrupación o clustering. Por otra parte, la disminución del número de variables es a lo que se le denomina reducción de dimensión de los datos y es comúnmente el primer paso en el desarrollo de métodos de aprendizaje supervisados. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.5.4. Exploración y visualización.

La exploración de los datos es de las primeras etapas del proceso de interacción con los mismos y está orientada a la comprensión desde un panorama global así como a la detección de valores inusuales en ellos. La exploración es utilizada en los procesos de minería para la limpieza y la manipulación de los datos, también en el análisis y descubrimiento visual. Los métodos de

exploración de datos incluyen la revisión de resúmenes desde las perspectivas gráficas y numéricas, pretenden mirar cada variable por separado y buscar las relaciones entre ellas, el propósito es descubrir excepciones o patrones. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

Por otra parte, la exploración mediante la creación de paneles o gráficos se denomina exploración grafica o análisis visual, para las variables de tipo numérico se suelen usar diagramas de cajas o histogramas con el objetivo de aprender sobre los valores de distribución de los mismos. En la detección de valores atípicos para encontrar información relevante en el análisis se categorizan las variables y se utilizan diagramas de barras para la representación, también se pueden observar como gráficos de dispersión de pares de variables numéricas con el fin de aprender las posibles relaciones, (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.6 Técnicas en modelos descriptivos

Los modelos de tipo descriptivo hacen referencia a la obtención de una clase (clasificación), comprador o no comprador, para tener más precisión y diferenciar con los modelos descriptivos, en la clasificación se busca obtener una clase mientras que en la predicción se desea conocer el valor continuo de una variable. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.6.1. Análisis de Clúster.

El análisis de clúster se usa para formar grupos de observaciones similares basados en diferentes medidas hechas de esas observaciones, la idea es caracterizar el clúster de manera que pueda ser útil para los objetivos del análisis. Este análisis se aplica en múltiples áreas como: Astronomía, medicina, química, educación, sicología entre otras. (Pei et al., 2012).

El método de clustering es una de las técnicas de aprendizaje no supervisado más populares donde el objetivo es segmentar los datos en grupos homogéneos de observación con el propósito de generar una idea, además sirven para mejorar el rendimiento de métodos supervisados, el modelamiento de cada clúster se aplica por separado en lugar de todo el conjunto de datos. El clustering es ampliamente utilizado en diversas aplicaciones de negocios como el mercadeo personalizado y el análisis de la industria. Los dos enfoques más populares del clustering son el clúster jerárquico y el K-means clustering, aunque existen otra técnicas como las redes neuronales auto-organizadas (SOM), también conocidas como redes de Kohonen. (Pei et al., 2012).

2.6.1.1. Clustering jerárquico.

En el clustering jerárquico las observaciones son agrupadas secuencialmente para crear clúster basados en la distancia entre las observaciones y la distancia entre los clúster y así producir una visualización jerárquica útil del proceso y de los resultados del agrupamiento llamado dendograma (tipo de representación gráfica de datos en forma de árbol). (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.6.1.2. Clustering K-means.

Otro tipo de clustering usado es el k-means, el cual es ampliamente utilizado en aplicaciones con grandes conjuntos de datos, en este método las observaciones son organizadas desde una perspectiva de conjuntos de clúster teniendo en cuenta la distancia de cada grupo (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.6.2. Redes neuronales auto-organizadas (SOM) o redes Kohonen.

Las redes auto-organizadas o redes neuronales SOM se caracterizan por tener una arquitectura compuesta por dos capas de neuronas, la capa inicial o de entrada está formada por cierto número de neuronas, una por cada variable de entrada, su función principal es la de recibir la información de entrada y hacer la transmisión de esta a la capa siguiente o capa de salida; en la capa de salida, se realiza el procesamiento de la información y la formación de mapas de rasgos, por lo general en estos mapas se organizan de forma bidimensional con el objeto de dar claridad al usuario en los resultados. (Marín, 1982).

2.7. Técnicas en modelos predictivos

A continuación se describen las técnicas relacionadas y que se aplican con mayor aceptabilidad en el uso de modelos de tipo predictivos, como las técnicas orientadas a clasificación entre otras. Los métodos de clasificación, regresión, reglas de asociación, filtrado colaborativo, constituyen la base de los métodos analíticos empleados en el análisis predictivo, sin embargo, en algunos casos el análisis predictivo incluye también métodos de identificación de patrones de datos como es el caso del clustering. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.7.1. Análisis predictivo.

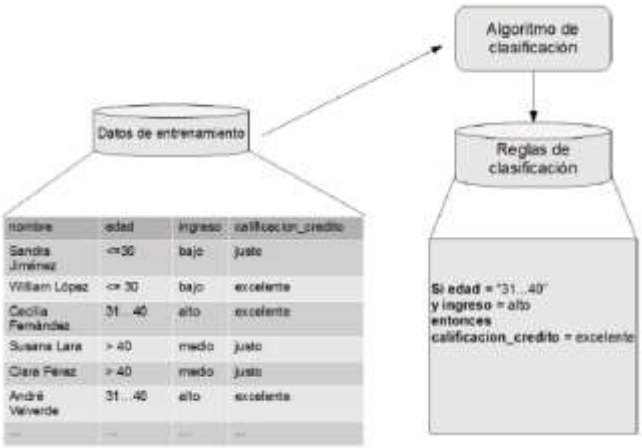
Los procesos de predicción son en esencia similares a los de clasificación con la diferencia que en la predicción se busca obtener el valor numérico de las variables (predicción continua), por ejemplo la cantidad de compras que un cliente realiza en un determinado centro comercial. En algunos casos la literatura utiliza los términos de regresión y estimación para referirse a la predicción del valor continuo de una variable; la predicción numérica es la tarea de predecir valores continuos. La predicción en variables de tipo discreto hace referencia a los valores donde la representación no posee un orden implícito, son modelos construidos para predecir etiquetas de variables categóricas (predicción discreta) como “si” o “no”, “tratamiento 1”, “tratamiento 2”, “seguro” o “de riesgo”. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.7.2. Clasificación.

Las tareas de clasificación dentro de la minería de datos son quizá de las más básicas en cuanto a análisis de datos se refiere, por ejemplo, el destinatario de una determinada oferta de una compañía puede responder a la misma o no responder, un cliente que solicita un crédito a una entidad bancaria puede pagar a tiempo, demorar en el pago o no pagar declarándose en quiebra, una

transacción con tarjeta de crédito puede llegar a ser fraudulenta o puede ser normal. Las tareas de minería de datos más comunes son aquellas en donde no se conoce la clasificación o esta ocurrirá a futuro, el objetivo es predecir que clasificación es o cual pudiera ser. Así mismo datos similares son utilizados en el caso de que la clasificación es conocida y son aprovechados en el desarrollo de reglas, las cuales posteriormente son aplicadas a los datos en donde la clasificación es desconocida. (Galit Shmueli, Peter C. Bruce, Mia L, 2014). En la figura 2 se puede apreciar un esquema de clasificación.

Figura. 2.Esquema de clasificación.



Fuente: (José Solano Rojas, 2010).

2.7.2.1. Regresión logística.

La regresión logística es un método bastante popular y poderoso dentro de los métodos de clasificación que como en la regresión lineal depende de una ecuación matemática específica que relaciona los predictores o entradas con la salida. El usuario debe especificar el predictor para incluir su forma; esto significa, que cada pequeño conjunto de datos puede ser usado para la construcción de un clasificador de regresión logística una vez que el modelo es estimado, es un clasificador económico y rápido computacionalmente incluso con muestras grandes de nuevas observaciones. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

La regresión logística extiende las ideas de la regresión lineal para situaciones en donde la variable dependiente es de tipo "categoría", se puede pensar en una variable categoría como la división de las observaciones al interior de las clases. Este método puede ser utilizado para clasificar una nueva observación donde la clase es desconocida basado en el valor de la variable predictor. También puede ser usada en datos donde la clase es conocida para encontrar factores distintivos entre observaciones en diferentes clases en términos de la variable predictor o el perfil predictor. La regresión logística es usada en aplicaciones como la clasificación de consumidores que regresan y no regresan, encontrar factores que diferencien los mejores ejecutivos entre hombre y mujer y predecir la aprobación o no de un préstamo basado en los registros de crédito. La idea principal bajo la regresión logística es en lugar de usar una variable dependiente se usa una función de esta

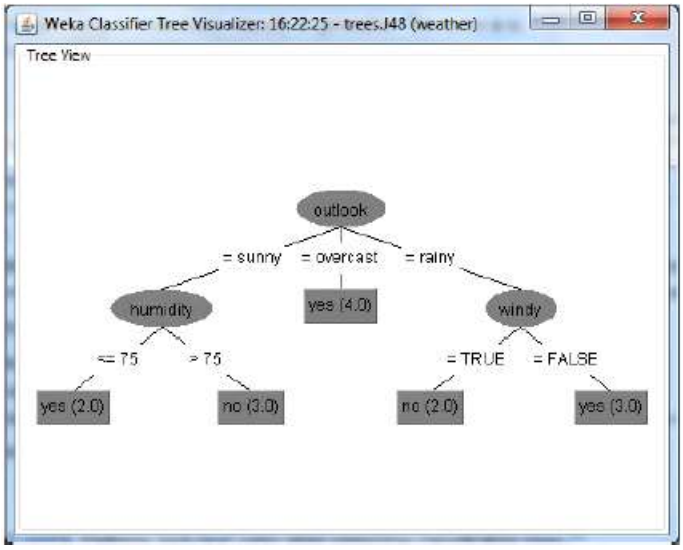
la cual es llamada “logit”, este a la vez puede ser modelado como una función lineal de los predictores, cuando el logit ha sido predicho puede ser asignado a una probabilidad. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.7.2.2. Clasificación por inducción de árboles de decisión.

La clasificación por inducción de árbol de decisión es el aprendizaje de árboles de decisión de clases etiquetadas de tuplas (lista ordenada de elementos) de entrenamiento, según (Pei et al., 2012) el uso de árboles de decisión en tareas de clasificación se presenta dada una tupla X, para la cual se asocia una etiqueta de clase desconocida, los valores de los atributos de la tupla se prueban contra los del árbol de decisión. Una ruta se traza desde la raíz a un nodo hoja, que contiene la predicción de clase para esa tupla. Los árboles de decisión se pueden convertir fácilmente en reglas de clasificación.

Un aspecto fundamental al momento de la construcción y el trabajo con técnicas de árbol de decisión es la poda de los mismos, ya que durante el entrenamiento muchas ramas reflejan anomalías debido a valores atípicos y ruido en los datos, existen métodos de poda que abordan el problema de sobreajuste de los datos, los cuales tratan de eliminar las ramas menos confiables haciendo uso de mediciones estadísticas, una árbol podado es más eficiente, más pequeño y más confiable que uno que no ha sido podado, por lo tanto, son menos complejos y más fáciles de entender, en cuanto a las tareas de clasificación son más rápidos y mejores para clasificar correctamente los datos de prueba independientes, es decir, las tuplas no vistas previamente. (Pei et al., 2012). Lo anterior se ilustra en la figura 3.

Figura. 3. Clasificación por arboles de inducción.



Fuente: (José Solano Rojas, 2010).

Generalmente las tecinas de poda tiene dos encauces principales: pre-poda y post-poda; en la pre-poda se busca la poda o detención temprana de la construcción del árbol, es decir al interrumpir el

crecimiento, el nodo se convierte en una hoja, la hoja puede contener la clase más frecuente entre las tuplas del subconjunto o la distribución de probabilidad de esas tuplas. En la post-poda se eliminan los sub-arboles de un árbol totalmente construido, de esta manera el subárbol se elimina de un nodo especifico y se reemplaza por una hoja la cual esta etiquetada con la clase más frecuente del subárbol eliminado. (Pei et al., 2012).

2.7.3. Clasificación Bayesiana.

Es importante para entender el método del clasificador de "Naive Bayes" conocer el clasificador Bayesiano completo, el principio básico en el que se fundamenta es sencillo y se estructura en 3 pasos:

- Encontrar todos los registros de entrenamiento con el mismo perfil de predictor, es decir los registros que tengan el mismo valor de predictor.
- Determinar a qué clase pertenecen los registros y cual prevalece en mayor grado.
- Asignar la clase que prevalece al nuevo registro.

Es recomendable ajustar el método para que responda a cuál es la probabilidad de pertenecer a una clase de interés o cuál es la clase más probable, mediante la probabilidad de clase es posible el uso de un deslizamiento de corte o límite para el proceso de clasificar un registro dentro de una clase particular, aunque la clase no sea la más probable. Este método es utilizado cuando hay una clase de interés que se desea identificar y la disposición para identificar diferentes registros como pertenecientes a dicha clase. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.7.3.1. Clasificador ingenuo de Bayes "Naive Bayes".

Con el objeto de dar solución a dificultades prácticas presentes en el método de Bayes completo se plantea el método de "Naive Bayes" en el cual se debe modificar el criterio en cuanto a la restricción del cálculo de probabilidad del registro, no necesariamente debe coincidir exactamente con el registro que va a ser clasificado, de hecho es posible usar un conjunto de datos de entrada, la modificación incluye cinco pasos:

- Para una clase particular se estima la probabilidad condicional individual para cada predictor, estas son las probabilidades de que el valor del predictor en el registro sea clasificado dentro de la clase particular.
- Multiplicar la proporción del registro en la clase establecida por el producto de las probabilidades condicionales.
- Repetir los paso anteriores en cada una de las clases.
- Para cada clase estimar una probabilidad de clase tomando el valor calculado en el paso 2 y dividirlo en la suma de los valores de todas las clases.
- Asignar el registro para la clase con la probabilidad estimada más alta del conjunto de valores de predictores.

Este clasificador es ideal en cuanto a simplicidad, eficiencia computacional, razonable en cuanto a rendimiento en la clasificación, hábil para manejar directamente variables categóricas. Su más eficiente respuesta se da cuando existe un gran número de predictores, sin embargo, esta es también parte de las desventajas del método ya que se requiere de una gran cantidad de datos disponibles para obtener muy buenos resultados. En la figura 4 se puede apreciar un ejemplo de ejecución del algoritmo Bayes Naive en la herramienta JPM, aplicado a un ejemplo de selección de características y la salida del sistema para 10 compañías de prueba. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

Figura. 4. Ejemplo de aplicación del método "Naive Bayes".

Fuente: (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.7.3.2. Redes de creencias Bayesianas.

En términos prácticos existen dependencias entre las variables ante lo cual el clasificador ingenuo de Bayes no es totalmente preciso, para esto el método de redes de creencia o confianza Bayesiana especifican la relación de distribuciones de probabilidad condicional conjunta, de esta manera permiten que las independencias condicionales de clase se definan entre subconjuntos de variables, así proporcionan un modelo gráfico de relaciones causales, sobre las cuales se puede aprender. Las redes de creencias bayesianas capacitadas se pueden usar para la clasificación, también se conocen como redes probabilísticas, están definidas por dos componentes: un conjunto de tablas de condicionalidad y un grafo acíclico, en este, cada nodo representa una variable aleatoria que puede ser de tipo continuo o discreto y así mismo pueden corresponder a atributos de los datos o a variables ocultas con los que forman una relación. (Pei et al., 2012).

2.7.4. Clasificación basada en reglas.

En el caso de los clasificadores basados en reglas el aprendizaje está representado por un conjunto de reglas, estas son una buena forma de presentar la información o conocimiento, la estructura "Si – Entonces" que presenta la expresión se describe así: SI – (condición) – Entonces – (conclusión); el "si" de la parte izquierda se conoce como la regla de antecedente o la precondición, en esta, la condición consiste en uno o más atributos de prueba, el "entonces" al lado derecho es la regla

consecuente y son agregadas de manera lógica, estas contienen una clase predicción. (Pei et al., 2012).

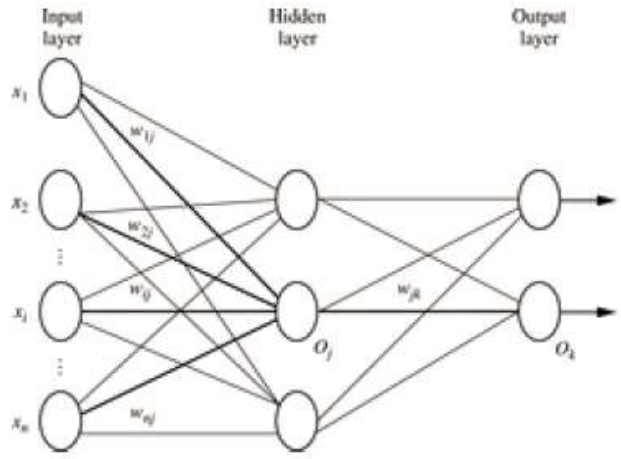
Si la condición en una regla antecedente resulta verdadera para una tupla dada, se puede afirmar que la regla está satisfecha y que además cubre la tupla, la cobertura de una regla está relacionada con el porcentaje de tuplas que cubre, lo que quiere decir que los valores de atributos que son verdaderos para el antecedente de la regla. Para determinar la precisión de una regla se busca las tuplas que cubre y el porcentaje que puede clasificar correctamente. (Pei et al., 2012).

Según (Pei et al., 2012). Existe un ordenamiento que puede ser basado en reglas o en clases, si el orden está basado en clases las mismas se dictaminan en orden de prevalecencía decreciente de acuerdo con su importancia, es decir las reglas de la clase más frecuente van primero y así sucesivamente. Adicionalmente pueden ser ordenadas basadas en el error de clasificación de costo por clase, en este caso las reglas no están ordenadas y no tienen por qué estarlo ya que todas predicen la misma clase así que no pueden existir conflictos de clase. Por otra parte, cuando el ordenamiento es basado en reglas, las mismas son organizadas dentro de una gran lista de probabilidad de acuerdo con alguna medida de calidad que puede ser el tamaño, la cobertura o la precisión, también pueden ser ordenadas de acuerdo a la observación de un experto en el dominio. Cuando se usa ordenamiento de reglas, el conjunto de reglas se conoce como lista de decisiones.

2.7.5. Clasificación basada en retro-propagación.

La retro-propagación refiere a un algoritmo de aprendizaje basado en redes neuronales, si bien existen muchos tipos de redes neuronales y algoritmos de redes neuronales, el algoritmo más conocido en cuanto a retro-propagación es la red neuronal multicapa de retroalimentación, esta red aprende de manera iterativa mediante un conjunto de pesos para la predicción de la etiqueta de clase; está estructurada en tres etapas, una capa de entrada, una o más capas ocultas y una capa de salida. (Pei et al., 2012).

Figura. 5. Red neuronal de alimentación multicapa.



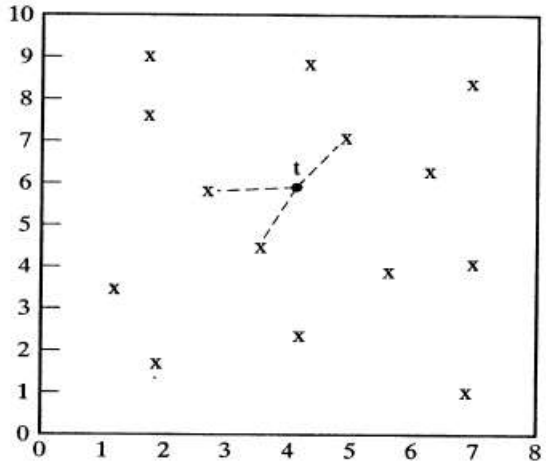
Fuente: (Pei et al., 2012).

Cada una de las capas de la red está compuesta por unidades, como se puede observar en la figura 5. Las entradas a la red son atributos medidos de cada tupla de entrenamiento, estas alimentan la capa de entrada de la red y a su vez pasan y alimentan la siguiente capa que es conocida como capa oculta o desconocida, las salidas de la capa oculta pueden ser entradas a otra capa oculta con lo cual el número de capas ocultas es arbitrario, las salidas ponderadas de la última de las capas ocultas se convierten en las unidades de entradas para la capa de salida, la cual es la encargada de la visualización de la predicción para las tuplas dadas. "*Las redes neuronales de retroalimentación multicapa pueden modelar la predicción de clase como una combinación no lineal de las entradas. Desde un punto de vista estadístico, realizan una regresión no lineal*". (Pei et al., 2012).

2.7.6. Clasificación supervisada. K-vecino más cercano (K-NN).

Este algoritmo hace referencia al vecino o vecindad más cercana, puede ser usado para tareas de clasificación como el resultado de una categoría y de predicción como un resultado numérico. Para predecir o clasificar un nuevo registro el método se basa en registros similares de los datos de entrenamiento, los vecinos son usados para derivar una clasificación o predicción de un nuevo registro, por votación para clasificación y por promedio para predicción. K-NN es un método altamente automatizado en el manejo de datos que presenta ventajas y debilidades en términos de rendimiento y consideraciones prácticas como el tiempo computacional. (Pei et al., 2012). Los métodos K-NN principalmente se basan en identificar los k registros en los conjuntos de datos de entrenamiento que son similares a un nuevo registro que se desea clasificar. Este método es no paramétrico ya que no involucra la estimación de parámetros en la función asumida tal como en la forma de la regresión lineal. Un ejemplo de aplicación de un algoritmo K-NN se puede observar en la figura 6.

Figura. 6. Clasificación usando K-NN.



Fuente: (José Solano Rojas, 2010).

El método K-NN para una respuesta numérica puede ser leído como predicción de valores continuos. El primer paso para determinar la distancia a la que permanece un vecino no cambia en comparación con el de respuesta de categorías, el segundo paso donde la mayoría de votos de los vecinos son utilizados para determinar clases es modificado de forma tal que se pueda tomar el valor promedio de respuesta de los K-NN para determinar la predicción, por lo general este es un promedio ponderado del peso decreciendo y la distancia aumentando del punto al cual se requiere la predicción. Otra de las modificaciones está en la medida del error para determinar el mejor K, en este caso se usan métricas de error cuadrático medio y error promedio absoluto. (Pei et al., 2012).

2.7.7. Máquina de soporte vectorial (SVMs).

En relación a esta técnica y según (Pei et al., 2012) la máquina de soporte vectorial es un método para clasificación lineal y no lineal, un SVMs es un algoritmo que funciona utilizando un mapeo no lineal para transformar datos de entrenamiento originales en una dimensión más alta, en esta se busca un híper plano de separación optimo no lineal que separa las tuplas de una clase de otra. El SVMs encuentra este híper-plano usando vectores de soporte y márgenes. (Pei et al., 2012).

Aunque el tiempo de entrenamiento de los SVMs incluso para los algoritmos más rápidos es extremadamente lento tienen la característica de ser altamente precisos debido a que poseen la habilidad para modelar decisiones complejas no lineales, esta característica los hace menos propensos al sobreajuste que otros métodos. Los vectores de soporte encontrados ofrecen una descripción compacta del modelo aprendido. Los SVMs pueden ser utilizados en tareas de predicción numérica como en tareas de clasificación. (Pei et al., 2012).

Por otra parte, (Carmona Suárez, 2014) confirma la definición antes mencionada en cuanto al algoritmo SVMs. En tanto en (Resendiz Trejo, 2006) presenta las SVMs como “ *un sistema para entrenar máquinas de aprendizaje lineal eficientemente tanto que para clasificación como para regresión se han encontrado muchas aplicaciones como clasificación de imágenes, reconocimiento de caracteres, detección de proteínas, clasificación de patrones, identificación de funciones*”.

2.8. Regresión

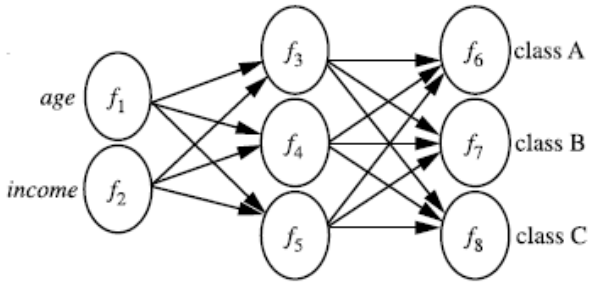
Las tareas de regresión son ampliamente utilizadas en el desarrollo de aplicaciones relacionadas con minería de datos, a continuación se describen algunas técnicas de uso común que son aplicadas en análisis de regresión.

2.8.1. Redes neuronales.

Las redes neuronales son un método flexible de manejo de datos usado para tareas de clasificación y predicción. A pesar de que en algunas ocasiones es considerado como una caja en blanco en cuanto a interpretación ha sido exitosa y posee alta exactitud en la predicción. Frecuentemente las

redes neuronales son llamadas redes neuronales artificiales, están basadas en un modelo de actividad biológica en el cerebro humano, en donde las neuronas están conectadas y aprenden de experiencias, la red neuronal imita la manera en la que los humanos aprenden. El aprendizaje y memoria de las redes neuronales se asemejan con el aprendizaje y memoria de los seres humanos y tienen la capacidad de generalizar de particularidades. La fuerza principal de la red neuronal está en su alto poder predictivo, su estructura soporta capturas de complejas relaciones entre los predictores y la respuesta, lo cual no es posible con otros modelos. (Galit Shmueli, Peter C. Bruce, Mia L, 2014). En la figura 7 se puede apreciar un esquema de una red neuronal.

Figura. 7. Red neuronal.



Fuente: (Han & Kamber, 2011).

El fundamento de las redes neuronales es combinar la información de entrada de una manera flexible y capturar relaciones complejas entre las variables de entrada y respuesta, por ejemplo en los modelos de regresión lineal las relaciones entre los predictores y la respuesta debe ser especificada directamente por el usuario lo que genera que la forma exacta de la relación sea muy compleja o sea generalmente desconocida, el modelo de regresión lineal trata de hacer diversas transformaciones a los predictores y la interacción entre los predictores pero la forma de la relación sigue siendo lineal. Por otro lado la red neuronal no requiere que el usuario especifique la forma correcta de la relación, de hecho la red aprende de cada relación de los datos, aun la regresión lineal y logística pueden ser consideradas como casos simples de redes neuronales que tienen solo capas de entradas y salidas y no capas ocultas. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

Existen numerosos y diferentes estudios sobre arquitecturas y aplicaciones de redes neuronales, entre las más exitosas aplicaciones en minería de datos han sido las redes de realimentación multicapa, estas son redes que poseen una capa de entrada compuesta por nodos llamados también neuronas que acepta valores de entrada y capas sucesivas que reciben entradas de las capas anteriores, los nodos de salida en cada capa son entradas para la capa siguiente hasta la última capa que es considerada como la capa de salida, las capas entre la entrada y la salida son conocidas como capas ocultas, una red de realimentación multicapa es una red completamente conectada con flujo unidireccional y sin ciclos.(Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.8.2. Regresión lineal múltiple.

Este método de regresión lineal múltiple es el más popular al momento de realizar predicción, se usa para ajustar a una relación entre una variable dependiente cuantitativa "Y" también conocida como resultado, objetivo o variable de respuesta y un conjunto de predictores también llamados variables de entradas o variables independientes X1, X2, aplicadas a una función matemática que aproxima la relación entre la entrada y la variable de salida.

$$Y = \beta o + \beta_1 X_1 + \beta_2 X_2 + \beta p Xp + \epsilon \text{ (Ecuación de regresión)}$$

En donde $\beta 0 \ldots . \beta p$ son coeficientes y ϵ hace referencia al ruido o la parte inexplicable, posteriormente los datos se utilizan para calcular los coeficientes y para cuantificar el ruido. Haciendo referencia a modelos predictivos los datos son usados para evaluar el rendimiento del modelo. Los modelos de regresión no solo significan la estimación de los coeficientes sino que también la selección de cuales variables de entrada y de qué forma serán incluidas. La regresión lineal múltiple es aplicable a numerosas situaciones de modelamiento predictivo, por ejemplo en la predicción del uso de tarjeta de crédito por parte de un consumidor con referencia a patrones de usos históricos y demográficos, prediciendo el tiempo de falla de la tarjeta basado en el uso de la misma. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.8.3. Árboles de regresión.

Los arboles de regresión son descritos como métodos flexibles de manejo de datos ya que pueden ser usados en clasificación, llamados arboles de clasificación o pueden ser usados en predicción, llamados arboles de regresión. Entre los métodos de manejo de datos estos son los más claros y fáciles de interpretar, son basados en observaciones de la separación al interior de subgrupos para crear divisores o predictores, estos crean reglas lógicas claras y de fácil entendimiento. Como otros métodos requieren de gran cantidad de datos para obtener buenos resultados, sin embargo, una vez construidos no son tan costosos de implementar incluso en muestras de grandes datos. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

El método de árbol puede ser usado en variables con respuesta de tipo numérico, en estos casos son conocidos como árbol de regresión y utilizados para tareas de predicción, la operación y el procedimiento ocurren de la misma manera que en el caso que se usan para clasificación pero la variable de salida en este caso es numérica. Se intentan crear muchos divisores y cada uno de estos debe ser comprobado mediante una prueba estadística en la que se usa un valor de registro para determinar el mejor punto de corte. Para regresión se aplica la prueba de la suma de los cuadrados y mide la diferencia media en los grupos para establecer el mejor punto de corte. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

En general los arboles de regresión producen fórmulas para hacer predicciones, las mismas pueden ser guardadas como registros en tablas. Estos métodos no son solo usados para clasificación y regresión también se utilizan en la selección de variables donde el predictor más importante muestra el mejor árbol. Los arboles requieren de relativamente poco esfuerzo de parte de los usuarios ya que no se requiere de transformación de variables, se debe tener presente que una

transformación monótona de variables puede generar los mismos arboles; El subconjunto de selección de variables es generado automáticamente ya que es parte de los divisores empleados en el proceso. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.9. Reglas de asociación y sistemas de recomendación en línea

El más claro ejemplo de reglas de asociación se presenta en las bases de datos de compras, en donde los datos de las transacciones de los clientes se prestan para realizar análisis de asociación entre los artículos comprados. Las reglas de asociación o análisis de afinidad como son también conocidas están diseñadas para encontrar patrones de asociación general entre artículos en grandes volúmenes de datos y pueden ser usadas de múltiples maneras dependiendo generalmente de donde se obtienen, si se hace referencia a datos de una tienda de consumo de alimentos, las reglas de asociación derivadas de estos datos se puede usar por ejemplo para establecer la ubicación más adecuada de los productos en los estantes. Si son derivadas de bases de datos hospitalarios podrían ayudar a encontrar síntomas recurrentes en pacientes y predecir síntomas futuros. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

Los sistemas de recomendaciones en línea como los que son usados por Amazon o Netflix usan un método de filtrado colaborativo que toma las preferencias individuales de los usuarios, las compras, calificación, navegación y otros aspectos que se puedan medir. Estos sistemas en caso contrario a las reglas de asociación que generan reglas generales para toda una población forman una asociación de "que va con que" de manera individual para cada usuario con el objetivo de entregar al mismo una recomendación personalizada y con una amplia gama de preferencias. (Galit Shmueli, Peter C. Bruce, Mia L, 2014).

2.10. Inteligencia de negocios (BI)

Según (Peña, 2006) "La Inteligencia de Negocios es el término que procura caracterizar una amplia variedad de tecnologías, plataformas de software, especificaciones de aplicaciones y procesos." Mediante esta premisa, el principal objeto de la inteligencia de negocios es lograr ventajas competitivas y contribuir a la toma de decisiones para mejorar el desempeño de la empresa. En (Shmueli, Patel, & Bruce, 2007; Vercellis, 2009), citan en conjunto la minería de datos y la inteligencia de negocios en procesos de optimización para la toma de decisiones acertadas y rápidas haciendo uso del saber concentrado de los procesos de la minería y la administración del mismo, teniendo como punto común la optimización del conocimiento y el soporte en las bases de datos.

Algunas comunidades académicas referencian la inteligencia de negocios como aplicaciones de tecnologías de información enfocadas a resolver problemas de negocios con grandes volúmenes de datos; comunidades profesionales la referencian directamente como inteligencia de negocios o como análisis de negocios (Chen et al, 2012). Por otro lado, según (Loshin, 2013) la define como: "los procesos, tecnologías y herramientas necesarios para convertir datos en información, información en conocimiento y conocimiento en planes que impulsan la acción comercial rentable.

La inteligencia de negocios abarca almacenamiento de datos, negocios, herramientas analíticas y gestión de contenido / conocimiento.

Por otra parte, desde el enfoque de (Cano, 2007), en su libro "Bussines Intelliigence: competir con información" pag 19, describe la inteligencia de negocios y las áreas con las que se relaciona de la siguiente manera:

> "BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un datawarehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios. Las áreas incluyen clientes, proveedores, productos, servicios y competidores."

Capítulo 3. Estado del arte

A continuación se presenta el estado del arte de la minería de datos aplicada a la inteligencia de negocios. Como criterios para la selección de aportes se decidió que se tomarían en cuenta aquellas contribuciones que eran aportes de la minería de datos a la inteligencia de negocios tomando como referencia la estructuración o áreas que cita (Cano, 2007), haciendo referencia al glosario de términos de (Garnet, 2006), es decir aquellos artículos que presentaron relación con: clientes, proveedores, productos, servicios y competidores. Se estructura el desarrollo del estado del arte en diferentes categorías teniendo en cuenta la información recopilada y producto del proceso de análisis, estableciéndose de la siguiente forma: fuentes de datos; a su vez está dividida en: datos de tipo estructurados y semi o no estructurados; selección de datos, exploración y visualización, relación de aportes de técnicas relacionadas con modelos de tipo descriptivo, incluyendo en esta categoría Clustering y subdivido en categorías donde se aplican técnicas relacionadas con los objetivos de este tipo de modelos. Continuando con modelos de tipo predictivos, a su vez enmarcados en subcategorías, estas se relacionan con clasificación y regresión. Finalizando con los aportes concernientes a reglas de asociación. Se presentan los aportes más significativos relacionados por los autores en cada una de las anteriores categorías y se resaltan los modelos, métodos o técnicas a los que se hace referencia en cada sección.

3.1. Información recopilada

En la figura 8, se describe de manera global la recopilación de aportes de los autores en orden cronológico desde el año 2010 hasta el año 2016. Se relaciona el número de publicaciones realizadas en cada año por parte de los autores y que posterior al análisis fueron tomados en cuenta para el desarrollo del trabajo. Es posible apreciar un crecimiento en la cantidad de publicaciones realizadas entre el periodo comprendido desde el año 2010 hasta el año 2013, posteriormente se aprecia un leve decrecimiento en referencia al año 2014, pero en el año 2015 se evidencia un aumento significativo de los aportes, lo anterior constituye un interesante aspecto de provecho en estas áreas de conocimiento y de apropiación por parte de la comunidad académica y científica.

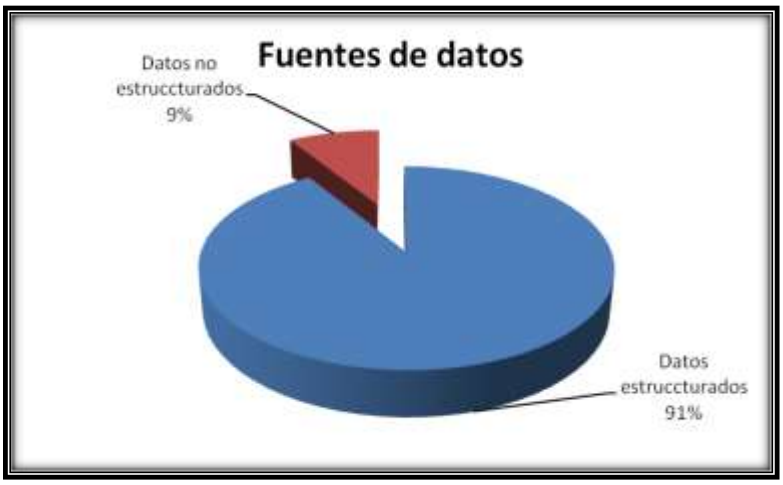
Figura. 8. Información recopilada.

Fuente: elaborada por el autor.

3.2. Fuentes de datos en minería de datos

Las fuentes de datos en los procesos de minería de datos son diversas, dependiendo del fin del proyecto, se pueden usar una o varias fuentes y pueden ser catalogadas de dos tipos principalmente, para este caso de estudio se enmarcan dentro de datos estructurados y semi o no estructurados. Tomando como referencia la figura 9, el porcentaje de trabajos con fuente de datos de tipo semi-estructurados y no estructurados es sustancialmente bajo en comparación con los estructurados. A continuación se relacionan los aportes encontrados en la revisión de la literatura.

Figura. 9. Fuentes de datos en minería de datos.



Fuente: elaborada por el autor.

3.2.1. Datos estructurados.

Las bases de datos con estructura más típica que relaciona fuentes de datos estructurados son las bases de datos relacionales y transaccionales, generalmente se almacenan en formato de tabla con etiquetas respectivas de fácil apropiación por parte de las herramientas de minería de datos para su tratamiento y análisis. Dentro de las fuentes de datos estructurados se pueden destacar varias categorías que son frecuentemente utilizadas en los procesos de minería de datos, se relacionan a continuación y se describen los aportes más relevantes.

3.2.1.1. Bases de datos relacionales.

Este tipo de bases de datos son una serie de tablas, cada una de las cuales tiene asignado un nombre único, cada tabla consiste en un conjunto de columnas o campos que generalmente almacena un conjunto grande de registros, cada registro posee una clave y se suele describir mediante valores de atributos. Los aportes relacionados se describen a continuación.

Uno de los enfoques donde se puede apreciar el uso de bases de datos de tipo relacional aplicado a la gestión de conocimiento y orientado hacia la caracterización de perfiles de estudiantes con deserción Universitaria se presenta en el trabajo de (Pinzon Cadena, 2011) se desarrolló sobre una base de datos real tomada de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda. La base de datos manejo las siguientes etiquetas: estado civil, país de nacimiento, estrato, medio por el cual se enteró del programa y de la Universidad, idioma, documento de identificación, semestre, ciudad de domicilio, ciudad de domicilio del acudiente, departamento de domicilio del acudiente, sexo y edad. Las técnicas utilizadas para el análisis fueron Clúster no jerárquico y algoritmo k-means.

Son múltiples los enfoques de aplicación de bases de datos relacional, uno de las tareas más típicas de la minería de datos y de aplicación de sus herramientas y técnicas se encamina a la predicción de fraude financiero, el trabajo desarrollado en (Ravisankar, Ravi, Raghava Rao, & Bose, 2011) se centra en esta problemática, para este caso se trabajó sobre una base de datos real que estaba compuesta por un conjunto de datos tomados de 202 empresas que cotizan en diversas bolsas Chinas, con 35 ítems financieros relacionados. Se utilizaron algunas técnicas de minería de datos como: (MLFF) red multicapa de alimentación de red neuronal, (SVM) máquinas de soporte vectorial, programación genética (GP), método de grupo de manejo de datos (GMDH), regresión logística (LR) y red neural probabilística (PNN). Cada una de las técnicas antes mencionadas se probó en el conjunto de datos. Se obtuvieron resultados que muestran que PNN supera a todas las técnicas sin selección de características, mientras que en el caso de técnicas con selección de características las que sobresalieron fueron GP y PNN presentando mayores grados de precisión.

Por otra parte, la ejecución de trabajos sobre base de datos relacionales con un enfoque donde se analiza la predicción del riesgo en el comportamiento de compra con tarjeta de crédito se da en el trabajo de (Yan-li & Jia, 2012) en este, se utilizan bases de datos de tipo real constituidas de 100 tablas de hojas de datos, en bases de datos de tarjetas de crédito. Para el análisis se seleccionaron seis tablas: tabla de información personal, tabla de consumidor, tabla de información de tarjeta,

tabla de registro de transacción, tabla de historial de sobregiro, tabla de saldo de historial. Se aplican técnicas como redes neuronales y algoritmo k-means.

Dentro de las fuentes de datos que se aplican a bases de datos de tipo relacional como las mencionadas en los trabajos anteriores, es posible identificar modelos de tipo descriptivo y predictivo aplicados. Teniendo en cuenta las técnicas utilizadas para el desarrollo de los trabajos es observable que algunas de las más consolidadas y usadas tradicionalmente se ejecutan como es el caso de: Clustering y algoritmos K-means, regresión logística, máquina de soporte vectorial. Sin embargo, adicionalmente se presentan algunas variaciones de las técnicas tradicionales y algunas técnicas de vanguardia como lo es el caso de la red neuronal de alimentación multicapa, la programación genética, el método de grupo de manejo de datos y la red neuronal probabilística.

3.2.1.2. Bases de datos transaccionales.

En términos generales una base de datos de tipo transaccional corresponde a un archivo en el que cada registro almacenado representa una transacción, en un formato típico una transacción incluye un número único de identificación de la transacción así como una lista de los elementos que hacen parte de la misma. A continuación se relacionan aportes encontrados en la revisión de la literatura en esta categoría de datos estructurados.

Para los centros comerciales, páginas de comercio electrónico y de compra en línea es de vital importancia el análisis del comportamiento de los usuarios, es relevante conocer este tipo de acciones con el propósito de orientar planes y estrategias de ventas, en el trabajo de (Devi, Devi, Rani, & Rao, 2012) se presenta un sistema inteligente para la agrupación de comportamientos de usuarios utilizando algoritmo de agrupación de tipo clúster jerárquico acumulativo, el sistema permite mejorar la taxonomía de la información de usuarios basada en datos transaccionales, es público y almacena información acerca de productos y revisiones así como de los comportamiento de los usuarios.

Desde otra perspectiva, el sector de mercado minorista, se desarrolla un trabajo en el cual el objetivo es generar grupos de productos basado en la extensión de la cesta de mercado, se presenta en el trabajo de (Ríos & Videla–Cavieres, 2014), para ello se utilizan registros transaccionales de una cadena minorista en Chile, alrededor de medio billón de registros que se reunieron en un período de veinte meses, aproximadamente 2, 200,000 clientes y más de 42,000 SKU en todo el mundo. En la ejecución se aplica un enfoque novedoso que genera comunidades de productos mediante algoritmos de propagación de etiqueta como: COPRA (actualiza los coeficientes de pertenencia promediando los coeficientes de todos sus vecinos) y SLA (es un algoritmo general de hablante-oyente basado en el proceso de propagación de la información) y minería gráfica que facilita la visualización de los resultados.

Dentro de la misma rama de base de datos transaccional resalta un trabajo novedoso que se aplica a una base de datos transaccional de datos inciertos, definida como información que frecuentemente se encuentra incompleta o con cierto grado de incertidumbre, este trabajo se

presenta en (Leung, MacKinnon, & Tanbeer, 2014) y plantea un modelo de estructura de árbol compacta para capturar datos inciertos mediante técnicas de minería de datos como: algoritmo de crecimiento de patrón frecuente e incierto TPC-growth, UF-growth (se encargan de escanear los datos dos veces para construir la estructura del árbol), en conjunto con estructura de árbol de patrones frecuentes (TPC-tree) para capturar de forma eficiente los contenidos de datos inciertos.

Dentro del análisis de los trabajos desarrollados sobre bases de datos de tipo transaccional resulta oportuno resaltar algunas de las técnicas utilizadas en los aportes presentados como es el clustering de tipo jerárquico acumulativo, una variación de este tipo clúster. Además, avances novedosos en cuanto a enfoques como los algoritmos de propagación de etiqueta y el algoritmo de mineralización de crecimiento de patrón frecuente e incierto, aplicado a bases transaccionales de tipo incierto.

3.2.1.3. Almacenes de datos o Data Warehouses (DW).

Un almacén de datos es un repositorio de información que puede provenir de múltiples fuentes, dicha información es almacenada mediante un formato unificado, generalmente dicha información se encuentra todo en un mismo sitio. Es modelado usualmente por una estructura de base de datos multidimensional en donde cada dimensión corresponde a un atributo o a un conjunto de atributos.

Una de las herramientas más poderosas y utilizadas en los procesos de minería de datos para el almacenamiento y tratamiento de los datos se enmarca en el uso de almacenes de datos, su construcción y arquitectura son aspectos relevantes al momento de iniciar las tareas de pre procesamiento y análisis de los datos, en este sentido el trabajo de (Luki, Radenkovi, Despotovi-Zraki, Labus, & Bogdanovi, 2016) se centró en proponer una nueva metodología de construcción de almacenes de datos en sistemas de inteligencia de negocios de múltiples dimensiones para operadores de redes eléctricas, basado en la metodología de ciclo de vida dimensional de Kimball (es una metodología detallada para el diseño), desarrollando e implementando sistemas de inteligencia de negocios y almacén de datos que se proporcionan a los usuarios finales para apoyar estrategias de desarrollo e implementación de almacenes de datos.

3.2.1.4. Bases de datos orientadas a objetos.

Las bases de datos relacional-objeto o bases de datos orientadas a objetos asumen principalmente que cada entidad en la base de datos es considerada como un objeto, la construcción de una base de datos de este tipo se realiza mediante un modelo de datos relacional de objetos. Este modelo mejora el modelo relacional al proporcionar un tipo de datos rico para manejar objetos complejos y orientación a objetos, teniendo en cuenta que esta es una de las necesidades de la mayoría de aplicaciones.

En el análisis de la bibliografía no se encontraron trabajos relacionados con este tipo de bases de datos.

3.2.1.5. Bases de datos temporales, bases de datos de secuencias y bases de datos de series de tiempo.

En relación a una base de datos de tipo temporal es aquella que posee atributos que implican marcas de tiempo, cada una de estas con una semántica diferente, generalmente este tipo de base almacena datos relacionales con atributos concernidos con el tiempo. Por otra parte, típicamente una base de datos de secuencia almacena secuencias de eventos ordenados, con o sin una noción del tiempo. Así mismo, una base de datos de series de tiempo almacena secuencias de valores o eventos obtenidos sobre mediciones repetidas de tiempo.

Un ejemplo de trabajos desarrollados sobre el tipo de bases de datos mencionadas anteriormente aplicado a desastres agro-meteorológicos donde los resultados sirven como apoyo en el proceso de toma de decisiones de cultivadores se desarrolla en (Peng, Zhang, Tang, & Li, 2011). El objetivo fue predecir mediante un marco de gestión de incidentes las emergencias y eventos futuros basados en información de incidentes anteriores. Se usan datos reales, tomados del registro de 5 años de desastres de este tipo que ocurrieron en China desde 1997 hasta 2001, con cinco bases de datos con la misma estructura, cada una reúne ocho atributos para representar el daño a los cultivos en 31 provincias chinas causadas por cuatro tipos de desastre agro-meteorológicos. Se aplican técnicas de minería de datos como: minería de patrones frecuentes, reglas de asociación y análisis de clúster.

Los procesos de desarrollo de estrategias de ventas y toma de decisiones son de gran importancia dentro de los métodos de negocios, un trabajo enfocado a un problema común como es la clasificación de los precios de los productos en estanterías se desarrolla en (Nafari & Shahrabi, 2010). Mediante un nuevo enfoque que clasifica adecuadamente los precios de los productos en estanterías teniendo en cuenta el cambio dinámico que estos presentan. Los datos usados son reales, sobre bases de datos de tipo secuencial tomados de la venta al por menor de cadenas de supermercados en Irán, cada transacción registra: tiempo de compra, los artículos comprados, los precios, descuentos y cantidades. Para el análisis se aplican reglas de asociación multinivel con el objeto de encontrar las relaciones entre los productos con respecto a sus precios y una modificación del algoritmo RApriori-TdMl (algoritmo para el descubrimiento de reglas de asociación multinivel).

En los trabajos relacionados anteriormente sobre los tipos de bases de datos mencionadas es prudente evidenciar el uso de métodos tradicionales que aportan al desarrollo propio de las aplicaciones y que se enmarcan dentro de las técnicas de modelos predictivos y descriptivos como en el caso del análisis de clúster y minería de patrones frecuente. Además, resalta el uso de variaciones de técnicas de reglas de asociación multinivel en búsqueda de encontrar relaciones de interés.

3.2.1.6. Bases de datos espaciales y bases de datos espacio-temporales.

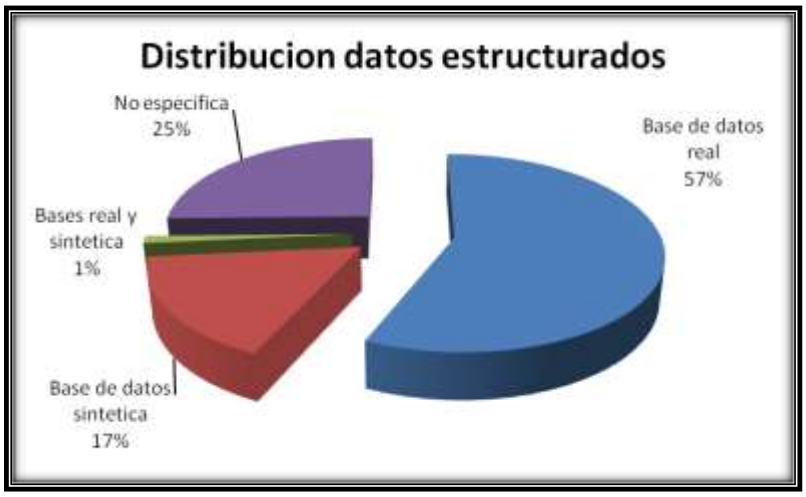
Las bases de datos de tipo espacial como su nombre lo indica contienen datos espaciales, estos pueden ser datos geográficos, datos de integración a muy grande escala, datos de diseño asistido

por computador, datos médicos y de imágenes satelitales. Este tipo de datos se representan bajo un formato específico llamado "raster" que consiste en un mapa de bits o de pixeles. Por otra parte una base de datos espacial que almacena objetos espaciales que son cambiantes con el tiempo se conoce como base de datos espaciotemporal con los que es posible por ejemplo, agrupar tendencias de objetos en movimiento.

En el análisis de la bibliografía no se encontraron trabajos relacionados con este tipo de bases de datos.

Un gran porcentaje de los artículos consultados en el actual trabajo presentan una fuente de datos de característica estructurada, la figura 10 muestra los porcentajes de aportes encontrados de fuente de datos estructurados y a su vez la distribución en cuatro categorías; los tomados de base de datos de tipo: real, sintética, real y sintéticas y los trabajos en donde la base de datos no está especificada dentro de datos utilizados. Con un mayor grado de participación de los trabajos desarrollados sobre bases de datos de tipo real.

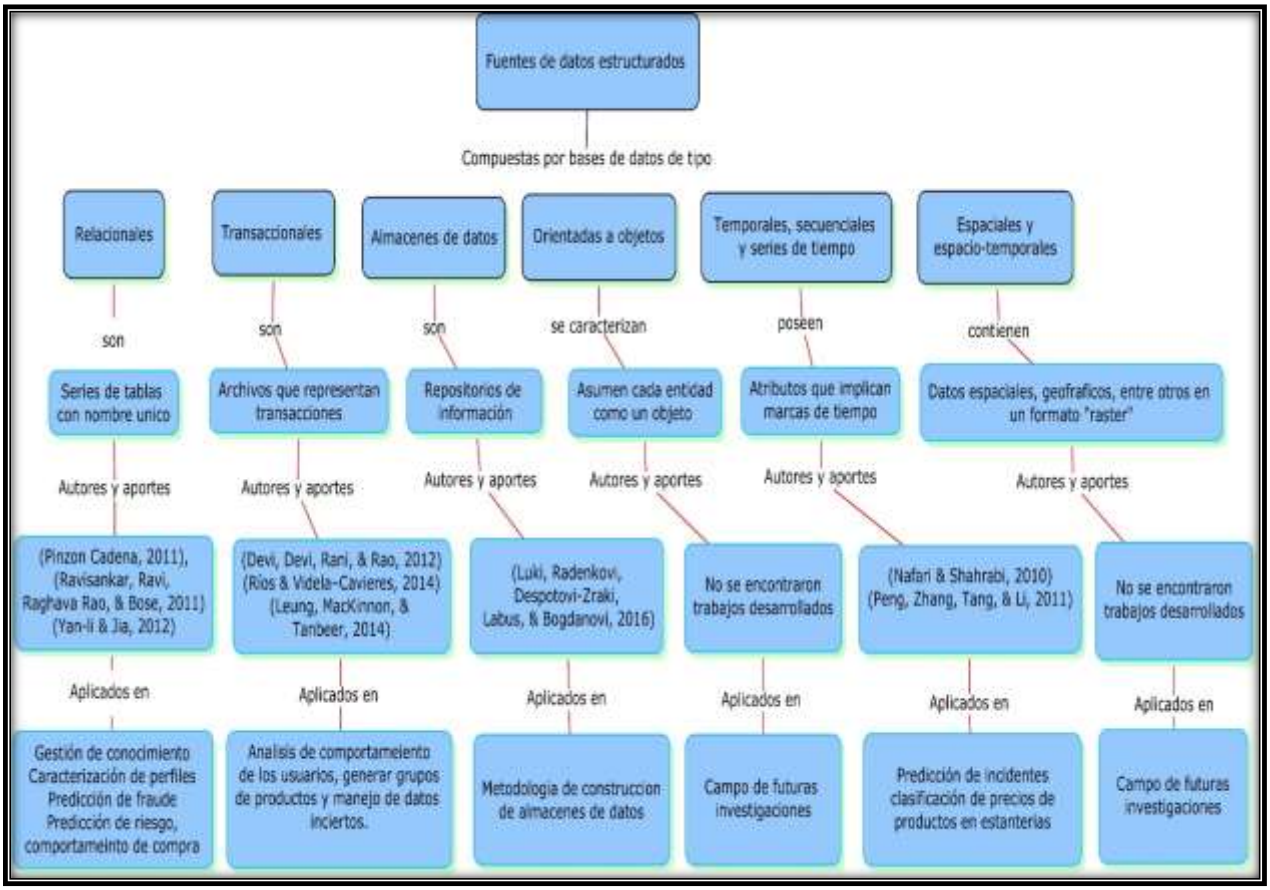
Figura. 10. Distribución bases de datos estructurados.



Fuente: elaborada por el autor.

Como aporte de esta sección en referencia a las fuentes de datos de tipo estructurados en la figura 11 se presenta un organizador grafico como resultado del análisis de la sección.

Figura. 11. Fuentes de datos estructurados.



Fuente: elaborada por el autor.

3.2.2. Datos semi-estructurados y no estructurados.

Si las fuentes de datos no son estructuradas, encasilladas en las enumeradas en la sección anterior, suelen encontrarse de dos tipos. Las fuentes no estructuradas, que no presentan un ordenamiento de los datos, es decir no están organizados en una estructura típica, sino que son obtenidos de bases de datos múltiples, generalmente son datos binarios que no están etiquetados o no tienen una organización donde exista un título que los identifique como pertenecientes a determinada categoría. Otro tipo de fuente, presentan cierto tipo de estructura, como el caso de las originadas en la web, en la que hay cierto tipo de etiquetado, pero en las que se requiere un tratamiento previo para extraer la información necesaria, este tipo de fuente se relaciona como semi-estructuradas.

Se pueden identificar algunas fuentes típicas de datos no estructurados tales como:

- Bases de datos de texto
- Bases de datos multimedia
- World Wide Web

3.2.2.1. Base de datos semi-estructurados.

Con objetivo principal de predecir síntomas y diagnóstico de enfermedades como la influenza para aplicarlas en sistemas de prevención desarrollados sobre plataformas web y redes sociales se presenta el trabajo realizado por (Corley, Cook, Mikler, & Singh, 2010), en este, los tipos de datos que se usaron fueron semi-estructurados, tomados de blogs con etiquetas como: título del blog, URL del blog, título del post, URL del post, fecha publicada (precisa en segundos), descripción, contenido completo codificado en HTML, etiquetas de asunto anotadas por autor e idioma. El análisis se apoyó en técnicas y herramientas como clúster jerárquico de 8 nodos y algoritmo de Girvan-Newman (algoritmo de búsqueda de comunidades, elimina los nodos con la mayor centralidad de inserción, identifica comunidades de interés), la minería de texto se utilizó para el monitoreo de las tendencias de la influencia en las redes sociales y plataformas web.

3.2.2.2. Bases de datos de texto.

Las bases de datos de textos son aquellas que contienen descripción de palabras para objetos, generalmente no son solo palabras claves sino descripciones de productos, informes de errores o mensajes de advertencia, estas bases de datos pueden estar altamente no estructuradas como es el caso de algunas páginas web, o pueden estar semi-estructuradas como los correos electrónicos o algunas páginas web. Las tareas de la minería de datos que usan fuentes de datos semi-estructurados o no estructurados se presentan en las descripciones de documentos de texto, identificación de palabras claves, el comportamiento de agrupación de objetos de texto entre otras. A continuación algunos aportes encontrados.

Trabajos de minería de datos enfocado a la inteligencia de negocios en donde se realiza manejo de bases de datos de texto en conjunto con minería de texto y otras técnicas se da por ejemplo en (Moro, Cortez, & Rita, 2015), para este caso el objetivo fue realizar una revisión de la literatura sobre inteligencia de negocios en el sector bancario. En el desarrollo del trabajo se aplicó minería de texto y el método de Dirichlet latente (método de distribución a priori, con bases en la estadística Bayesiana), el modelo de asignación de Dirichlet latente se utilizó para agrupar artículos en diferentes tópicos relevantes. Como resultado se pudo identificar las relaciones entre los términos y los temas que agrupan los artículos. Por otra parte en (Amarouche, Benbrahim, & Kassou, 2015) el estudio se centró en la clasificación de técnicas de minería de datos aplicadas a inteligencia competitiva y minería de producto de opinión. Se usaron técnicas de máquina de aprendizaje, procesamiento de lenguaje natural (PLN) y minería de opinión con el fin de clasificar un comentario como una opinión positiva o negativa y de esta manera apoyar la inteligencia competitiva.

3.2.2.3. World Wide Web.

Posterior al análisis de la bibliografía en cuanto a este tipo de fuente de datos se encontraron algunos aportes con origen de datos no estructurados o semi-estructurados especialmente sobre la World Wide Web, que se relacionan a continuación.

Es de relevancia para las empresas conocer la apreciación de sus clientes en muchos aspectos, específicamente para analizar si están siendo efectivas las estrategias de mercadeo e identificar cuáles son los factores de éxito y cuáles prácticas no le están siendo favorables. Un trabajo orientado en este sentido se realizó en (Thorleuchter & Van Den Poel, 2012), se planteó un modelo basado en conceptos semánticos latentes (un enfoque de minería de texto que se refiere a todos los conceptos con igual significado) aplicado sobre bases de datos reales, los datos fueron recolectados de páginas web de comercio electrónico usando métodos como tokenización y varios métodos de filtrado. Para el análisis se aplicaron patrones de información textual en la clasificación de texto de sitios web y la identificación de factores de éxito de empresas de comercio electrónico, la minería de texto se aplicó en conjunto con un modelo de regresión logística. Los resultados favorables en cuanto a la predicción del éxito de empresas de comercio electrónico apoyan los procesos de planificación empresarial.

Adicionalmente al aporte mencionado anteriormente, se han desarrollado y aplicado trabajos con el objetivo de cotejar las relaciones de revisores de consumidores en línea y comparar métodos de aprendizaje de minería de opinión a nivel de las característica de los algoritmos en diferentes sitios populares de comercio electrónico como Yahoo Shopping y Amazon.com se presenta en (L. Chen, Qi, & Wang, 2012). Se aplicaron técnicas de minería como reglas de asociación, minería de opinión mediante el modelo de Markov oculto lexicalizado (L-HMM), modelo de campos aleatorios condicionales (CRF), dos variantes de métodos basados en minería de reglas de asociación, es decir, minería de reglas de asociación (ASM) y ASM más reglas lingüísticas basadas en L-HMM en CRF.

Por otra parte, con el objetivo de extraer conceptos de escenarios de datos futuristas en documentos textuales mediante la integración de técnicas en escenarios de manejo de este tipo de datos en un novedoso campo de desarrollo, presentó el trabajo de (J. Kim, Han, Lee, & Park, 2016) el cual se desarrolló sobre la utilización de datos futuristas, una colección de opiniones orientadas al futuro, extraídas de las comunidades en línea. Para el análisis se aplicó la técnica de minería de reglas de asociación difusa (FARM), con el fin de identificar ponderaciones casuales en función de reglas, integrando los mapas cognitivos difusos (FCMs), minería de texto y el análisis semántico latente (LSA).

En el análisis de esta categoría es importante resaltar algunos métodos, modelos, tipos de minería novedosos como lo son: modelo de Markov oculto lexicalizado (L-HMM), modelo de campos aleatorios condicionales (CRF), métodos basados en minería de reglas de asociación, minería de reglas de asociación (ASM) y ASM más reglas lingüísticas basadas en L-HMM en CRF. Además un tipo de minería relacionado con reglas de asociación difusa (FARM) y el análisis semántico latente (LSA).

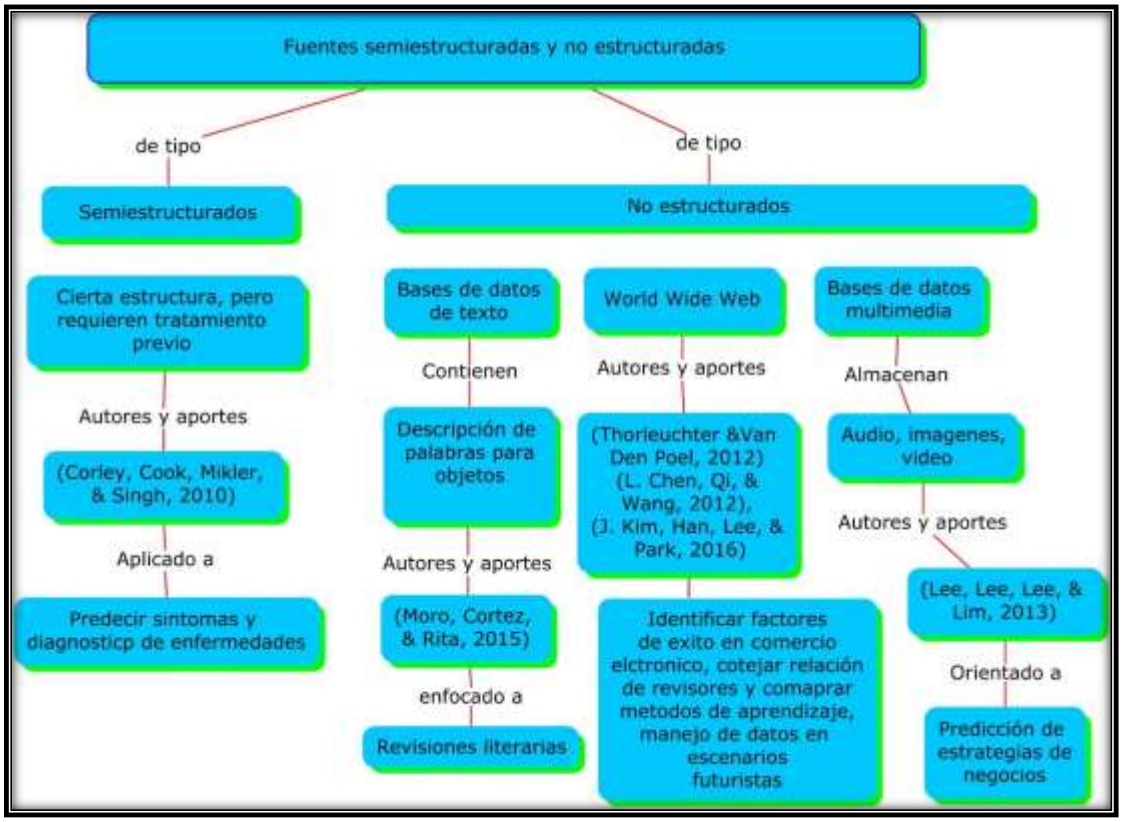
3.2.2.4. Bases de datos multimedia.

Las bases de datos multimedia como su nombre lo indica almacenan datos de tipo multimedia como: audio, video e imágenes; las aplicaciones con este tipo de datos suelen orientarse hacia

sistemas de correo de voz y sistemas de pedido por video. Las tareas de minería multimedia requieren la integración con métodos de minería de datos, especialmente para búsqueda y almacenamiento. Se relacionan a continuación algunos aportes afines encontrados en esta categoría posterior al análisis de la información recopilada:

Un trabajo donde se aborda el problema de la predicción de estrategias de negocios en empresas del sector industrial que permita simular la toma de decisión en datos no estructurados al considerar la incertidumbre sobre la variables y facilite el trabajo de los gerentes sobre escenarios hipotéticos se desarrolla en (Lee, Lee, Lee, & Lim, 2013) mediante un método de inferencia basado en agentes inteligentes para el manejo del tiempo y las relaciones dinámicas, se desarrolló sobre una base multimedia de datos real y el simulador MACOM (multi-agente basado en mapas cognitivos difusos). Se aplicaron técnicas como mapas cognoscitivos difusos (FCM), representados mediante un gráfico firmado que consta de nodos y bordes. En el experimento, la recopilación de información parcial se realizó a través de cuatro pasos: obtención de datos brutos, consenso, obtención de relaciones preliminares y cuantificación de las relaciones. Como resultado de la sección anterior se presenta el organizador grafico de la figura 12.

Figura. 12. Fuentes semi-estructuradas y no estructuradas.



Fuente: elaborada por el autor

3.3. Selección de datos

Dentro de la clasificación y análisis realizado en la bibliografía se encuentran algunos modelos, métodos y herramientas relacionadas con la selección de datos, constituyen el 16 % de los aportes en la bibliografía, algunos de estos son listados en la tabla 1 que se presenta a continuación.

Tabla 1. Selección de datos en minería de datos.

Modelos, métodos y herramientas de selección de datos	Autores
Modelo CRISP-DM (adaptación a los procesos de negocios)	Wegener & Rüping, 2010
Métodos de tokenización, filtrado colaborativo y descubrimiento de patrones de información textual	Thorleuchter & Van Den Poel, 2012)
Sistemas de soporte de decisiones (definición de requerimientos)	Demirkan & Delen, 2013
Metodología de procesamiento (Minería web) en privacidad de la información	Velásquez, 2013
Escala Likert (comprensión de opiniones y actitudes), modelo TTF (tecnología ajuste de tareas), ECM(modelo de espera confirmación)	Huang, Wu, & Chou, 2013
Patrones de datos intangibles o inciertos	Do, Bae, & Park, 2015
Nueva metodología de construcción de almacenes de datos.	Luki, Radenkovi, Despotovi-Zraki, Labus, & Bogdanovi, 2016

Fuente: elaborada por el autor.

En esta sección son de relevancia los procesos relacionados con la minería de texto y la selección de datos textuales; cuando se hace referencia a trabajos sobre minería de texto se habla del proceso de sacar información potencialmente útil de documentos y otras fuentes afines como los correos electrónicos, mediante la identificación de patrones dentro de los textos, estos pueden ser estructuras semánticas, usos de palabras, identificación de anuncios influyentes, identificación de palabras claves, revisión sistemáticas de literatura, entre otras.

Continuando con las consideraciones anteriormente mencionadas, los procesos de selección de datos textuales hacen referencia a datos tanto de tipo cualitativo como cuantitativo, generalmente en estos procesos intervienen gráficas y matrices de tipo descriptivo, explicativo y análisis de contenido, en el cual se busca el descubrimiento de significado o presencia de este en un documento.

A continuación se relacionan algunos de las contribuciones en cuanto a selección de datos, generalmente se aplican a fuentes de datos estructurados dada la mayor facilidad de manejo y obtención de resultados fáciles de interpretar por los usuarios.

Uno de los trabajos de selección de datos se orientan a la clasificación de texto en sitios web como en (Thorleuchter & Van Den Poel, 2012). Se recolectó información de páginas web de comercio electrónico mediante métodos de tokenización (en el análisis de información textual la palabra se considera como una unidad y el método se aplica para convertir términos en minúsculas y capitalizar el primer carácter) y varios métodos de filtrado (los métodos de filtrado, colaborativos

y de otros tipos, se caracterizan por ser técnicas usadas muy frecuentemente por los sistemas de recomendación, específicamente se busca la reducción de información no útil a la que con el auge de la tecnología y el internet se tiene disposición. Los tipos de métodos de filtrado pueden ser híbridos, demográficos, colaborativos y algunos basados en contenido, sirven principalmente para hacer predicciones automáticas sobre los intereses de los usuarios).

Desde otro punto de vista en (Zhang, Mukherjee, & Soetarman, 2013) presentó un método supervisado de extracción automática de frases claves (KEA) basado en un modelo de Naive Bayes y un método no supervisado extractor automático de conceptos (ACE) que analiza textos y HTML. El objetivo fue identificar de qué se está hablando en una página web relacionada con compras, la fuente de datos es de tipo real, tomados de 100 páginas web relacionadas con compras de sitios de marcas líderes en el mercado como Dell, HP y Canon. Se realiza una mejora al método ACE convirtiéndose en extractor de conceptos mejorado (ICE), por otra parte, la minería de opinión se aplica para realizar el análisis de sentimientos y emociones, esta aplicación se transforma en apoyo a los sistemas de toma de decisiones.

Resulta oportuno mencionar el trabajo desarrollado por (Do, Bae, & Park, 2015), se centró en encontrar patrones de datos intangibles e inciertos en procesos de desarrollo de productos. Aplica un método de minería analítica en línea (OLAM) basado en la técnica de análisis de datos de productos que examina y selecciona los datos de las bases de gestión de producto y los integra con el procesamiento analítico en línea (OLAP). Los datos se obtuvieron de la base de gestión de datos de productos (PDM).

Por otra parte, los almacenes de datos tienen un papel predominante en la selección de los mismos, en (Luki et al., 2016), se propuso una nueva metodología de construcción de almacenes de datos en sistemas de inteligencia de negocios de múltiples dimensiones para operadores de redes eléctricas, se aplicó la metodología de ciclo de vida dimensional de Kimball, es una metodología detallada para el diseño, desarrollando e implementando sistemas de inteligencia de negocios y almacén de datos.

De esta sección es importante acotar que existen técnicas tradicionales bien marcadas como lo son la regresión logística y el método de Bayes ingenuo, utilizadas como base para el desarrollo de nuevos avances, se evidencian también el uso de los almacenes de datos en los procesos de selección y métodos propios de selección y extracción como lo son el filtrado y tokenización. Además, los de los métodos de extracción de frases (KEA) y de conceptos claves (ACE), así como una mejora a este último, llamado (ICE).

3.4. Exploración y visualización

En la siguiente sección se describirán las técnicas y herramientas encontradas posteriores al análisis de los aportes en cuanto a exploración y visualización.

3.4.1. Exploración de datos.

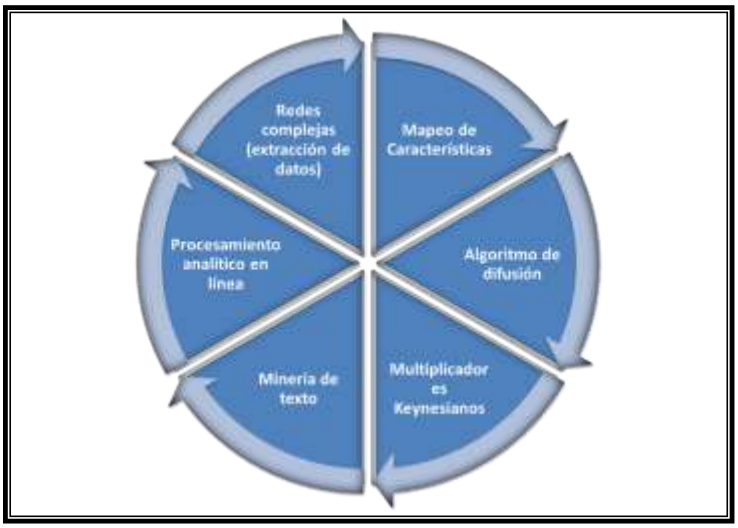
Generalmente las tareas de exploración de datos involucran diversas herramientas para el tratamiento de los mismos como la limpieza y transformación. A continuación se relacionan algunos de los aportes:

Las técnicas y herramientas más utilizadas para la exploración de datos en el desarrollo de los aportes se relacionan con mapeo de características, exploración de datos mediante redes complejas, procesamiento analítico, multiplicadores Keynesianos entre otros.

Uno de los trabajos enfocados en la exploración de datos se presentó en (Zhu, 2013), quien realizó mediante un novedoso algoritmo dinámico de difusión que trata de encontrar clústeres en la red social de confianza del usuario maximizando la tasa de cobertura de la red en cada ciclo, los nodos influyentes encontrados en un primer ciclo se usan como la fuente de difusión en el ciclo siguiente. Es decir, el algoritmo utiliza el pensamiento iterativo para seleccionar los nodos influyentes.

Desde una perspectiva teórica (Kopaneli, 2014) El trabajo se centró en descubrir la conexión entre el análisis económico financiero y la contabilidad. Para la exploración y el desarrollo se aplicaron modelos estadísticos que calculan los efectos de empleos a corto plazo, como los modelos multiplicadores keynesianos o multiplicadores del gasto público. En cuanto a la exploración de los datos las herramientas y metodologías encontradas se presentan en la figura 13.

Figura. 13. Exploración de datos.



Fuente: elaborada por el autor.

3.4.2. Visualización de los datos.

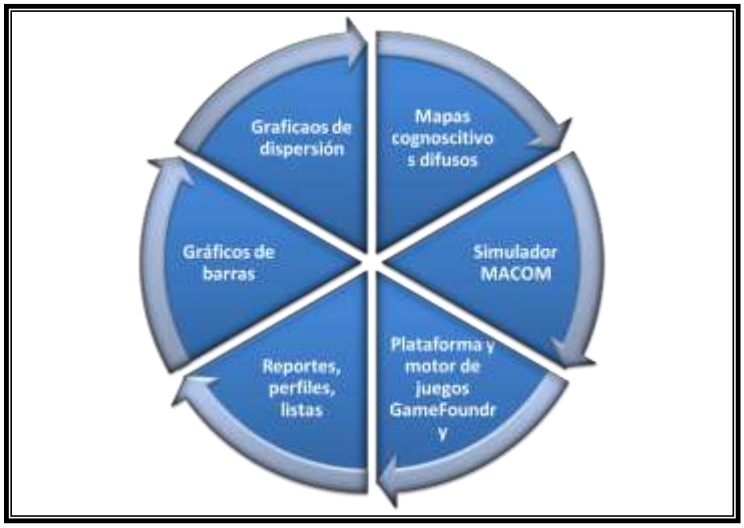
Las tareas de visualización permiten hacer más fácil la interpretación de los resultados de los procesos de minería de datos. A continuación se relacionan algunos de los aportes encontrados:

En cuanto a visualización las herramientas y técnicas más utilizadas se relacionan con gráficos de dispersión, diagramas de barras, mapas cognoscitivos, simuladores y herramientas de

visualización de reportes entre otras, algunos autores como (Popeangă & Lungu, 2012), (Aufaure, Chiky, Curé, Khrouf, & Kepeklian, 2016) desarrollaron trabajos sobre la base de soluciones de inteligencia de negocios en tiempo real (RTBI) en sectores de servicios públicos y gestión de transmisión de datos semánticos, estos sistemas ofrecen un alto de grado de visualización de la información, lo que permite una mejor comprensión de los resultados.

La visualización en los sistemas de RTBI requieren el desarrollo de una adecuada interfaz que permita la facilidad de comprensión de los resultados para la toma de decisiones por parte del usuario y que a su vez realimente el sistema de procesamiento, análisis y transformación de los datos, estas herramientas RTBI soportan los procesos de toma de decisiones y ofrecen una interfaz de usuario más amigable que las que presentan las técnicas de minería de datos por si solas. Para los estudios mencionados los datos fueron tomados del software "Oracle Business Intelligence for Utilities". Las herramientas de visualización encontradas posterior al análisis de la bibliografía se relacionan en la figura 14.

Figura. 14. Visualización de datos.



Fuente: elaborada por el autor.

3.5. Modelos descriptivos

En esta sección se relacionan los aportes encontrados en el análisis de la literatura concerniente a los objetivos propios de las técnicas que se aplican en el desarrollo de estudios y trabajos dentro de los modelos descriptivos, en el análisis se evidencio principalmente las técnicas de clustering y su sub categorías de este enfoque orientados a diferentes sectores. Además se relacionan aportes relevantes.

3.5.1. Clustering

Dentro de las tareas de clustering encontradas en la revisión de la literatura es posible resaltar ciertos enfoques como: descubrimiento de patrones sobre los precios de oferta en el mercado de publicidad, la gestión de conocimiento aplicado a la deserción universitaria, la segmentación de consumidores en comercio electrónico, la agrupación de comportamientos de búsqueda de los usuarios, la clasificación de revisores e identificación de preferencias de los usuarios, el descubrimiento de intereses de los consumidores y el soporte para los procesos de clasificación de reglas de inducción.

3.5.1.1. Clustering aplicado al descubrimiento de patrones, precios en mercados e intereses de los usuarios.

En cuanto al descubrimiento de patrones sobre precios de oferta en mercados de electricidad en (Eduardo & Vega, 2011) el propósito fue cuantificar hipótesis sobre el efecto de algunas variables de ofertas y precios, en particular, cuantificar la confianza sobre el efecto de las condiciones hidrológicas en las pujas y precios al contado en un mercado fuertemente dependiente de plantas hidroeléctricas, para esto se aplicaron técnicas de reglas de asociación difusa y se trabajó un algoritmo de agrupación que se desarrolló sobre una base de datos multidimensionales de precios de licitación, tomados de los 10 agentes generadores más grandes del mercado eléctrico colombiano. El algoritmo de aprendizaje no supervisado de clustering k-mean se aplica para la representación de funciones como puntos en un espacio multidimensional.

Desde otro punto de vista, se propuso un nuevo método para descubrir intereses de los consumidores en (Su & Chen, 2015) aquí se extraen categorías a través de flujos de clics registrados un sitios web de comercio electrónico, para esto se usó un algoritmo de clustering de líderes que se basa en las similitudes de los usuarios en términos de comportamiento de navegación. El análisis de conjunto aproximado permite que un usuario se asigne a más de un clúster, las categorías involucradas en cada grupo se ilustran mediante un grafo de colores de círculos múltiples donde el grafo representa los patrones de interés de los usuarios en el clúster correspondiente. Se recopilaron datos de: transmisión de clics de usuarios de PC, las URL solicitadas y las marcas de tiempo de las solicitudes, el conjunto de datos original se compone de aproximadamente tres millones de registros que fueron aleatoriamente seleccionados de los registros del servidor.

Las técnicas utilizadas en el análisis de esta categoría se enmarcan dentro de las siguientes: reglas de asociación en conjunto con algoritmos de agrupación, técnicas de clúster no supervisado como K-means. Además se plantea una mejora al algoritmo de clustering convencional haciendo un algoritmo de clúster líder enfocado en las similitudes de comportamiento.

3.5.1.2. Clustering orientado a la gestión de conocimiento y caracterización de perfiles.

Un enfoque hacia la caracterización de perfiles de estudiantes con deserción Universitaria se presenta en (Pinzon Cadena, 2011) el objetivo fue hacer gestión de conocimiento sobre la manera

de concentrarse en la información de las bases de datos empleadas en las instituciones académicas, para ello se utilizaron técnicas de Clúster no jerárquico y algoritmo k-means; sobre una base de datos real tomada de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda, donde se almacenan los registros de los estudiantes desde el ingreso a la Universidad. Se tuvieron en cuenta ciertas variables como: estado civil, país de nacimiento, estrato, medio por el cual se enteró del programa y de la Universidad, idioma, documento de identificación, semestre, ciudad de domicilio, ciudad de domicilio del acudiente, departamento de domicilio del acudiente, sexo y edad. Los resultados muestran que las causas identificadas de deserción de los estudiantes son: aspectos económicos, bajo rendimiento académico y embarazos a temprana edad teniendo en cuenta que el área de conocimiento de los programas a los cuales se aplicó el estudio está conformado en alto porcentaje por mujeres.

3.5.1.3. Clustering aplicado a la segmentación de consumidores, agrupación de comportamientos de búsqueda.

Uno de los enfoques del clustering se orienta la segmentación en línea de consumidores en comercio electrónico como en (Wu & Chou, 2011), en este trabajo cada cliente tiene una membrecía mixta y un puntaje asociado con cada clase latente, los clientes son asignados a múltiples clases latentes si sus puntajes de membrecía para estas clases son suficientemente similares. Para la segmentación se usaron múltiples categorías como: satisfacción con el servicio, uso de internet, comportamiento de compra y características demográficas. La técnica de minería usada fue un método de clúster suave con clases latentes mixtas derivada del modelo de asignación de Dirichlet latente. Los resultados de la aplicación de este método proporcionan a los gestores información útil y herramientas para la mejora de las relaciones con los clientes, así mismo mejora los procesos de segmentación que aplican métodos de clúster duro y presenta mayor calidad de agrupamiento.

Por otra parte el trabajo de (Devi et al., 2012) presenta un sistema inteligente para el clustering de comportamientos de usuarios con el fin de clasificar los comportamiento de exploración de los usuarios dada la heterogeneidad de las características de búsqueda. Se utilizó algoritmo de agrupación de tipo clúster de jerárquico acumulativo, el sistema permite mejorar la taxonomía de la información de usuarios basada en datos transaccionales, es público y almacena información acerca de productos y revisiones así como del comportamiento de los usuarios. Los datos son tomados de la web, o archivos de registro web logs tomados de servicios web como RapidMiner, Digg.com, Amazon, eBay.

De acuerdo con los razonamientos y aportes expuestos en esta categoría resulta oportuno resaltar la aplicación de métodos de clúster suave en conjunto con el uso de clases latentes mixtas así como el uso de un algoritmo de agrupación de tipo clúster de tipo jerárquico acumulativo que mejora la clasificación de información de los usuarios.

3.5.1.4. Clustering aplicado a identificación de preferencias de los usuarios y clasificación de revisores.

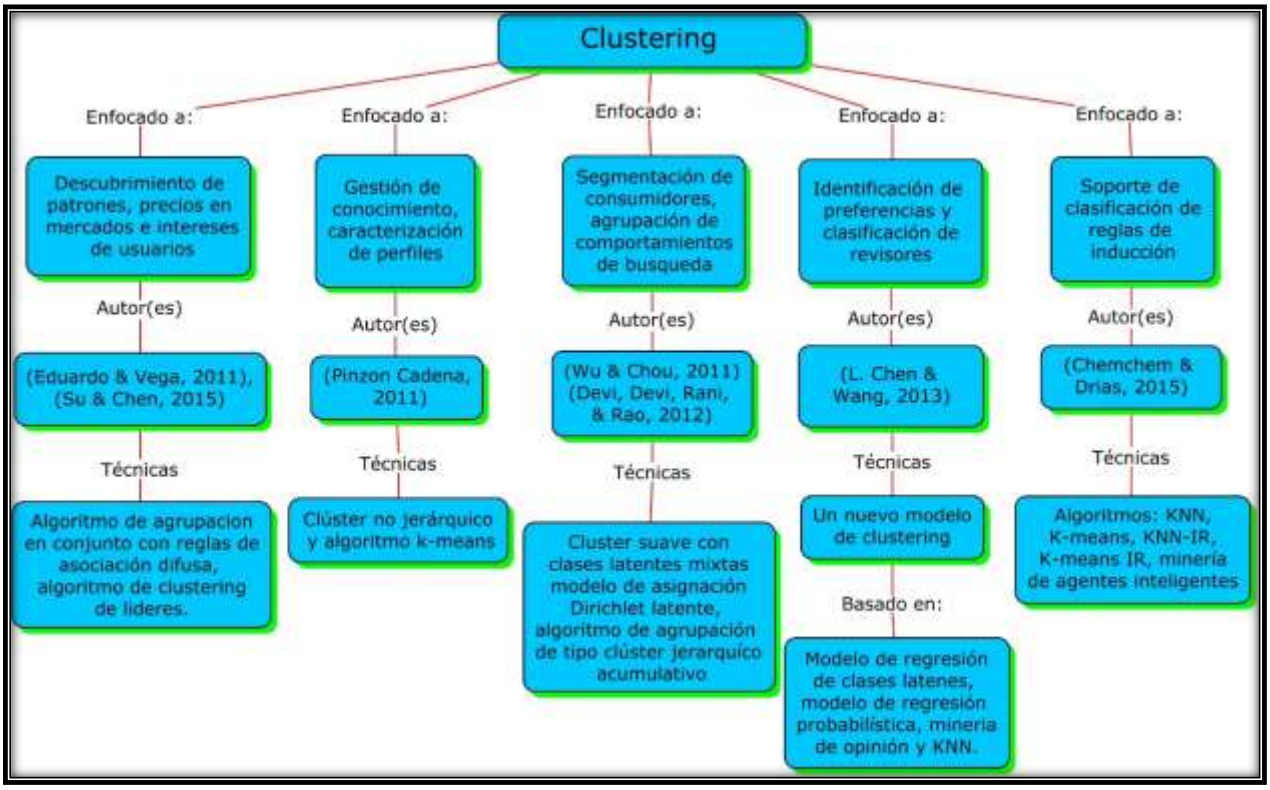
Un aporte relacionado con esta categoría se plasma en el trabajo de (L. Chen & Wang, 2013) en este se presenta un nuevo método de agrupamiento (clustering) para considerar características y calificaciones de revisores e identificar preferencias de usuarios, el método está basado en un conjunto de técnicas: el modelo de regresión de clase latente LCRM, un modelo de regresión probabilística (PRM), minería de opinión y KNN. Se utilizaron dos bases de datos del mundo real, un conjunto de datos de cámaras digitales y un conjunto de datos de portátiles que se rastrearon desde un sitio web comercial "www.buzzillions.com". Este método es capaz de considerar calificaciones generales y valores de opinión a nivel de características dadas por los revisores de los productos en los diferentes conjuntos de datos recolectados.

3.5.1.5. Clustering enfocado al soporte de clasificación de reglas de inducción.

La clasificación de reglas de inducción se constituye una de las tareas más relevantes dentro de las aplicaciones del clustering. Un nuevo enfoque de soporte para la agrupación y clasificación de reglas de inducción se propone en (Chemchem & Drias, 2015). Se utilizan para tal fin técnicas como: algoritmo K-NN, K-means, los algoritmos K-NN-IR y K-means-IR (modificados para el tratamiento de reglas de inducción), minería de agentes inteligentes, un nuevo enfoque de arquitectura de minería (MIA) Miner Intelligent Agent, este hace referencias a un agente cognitivo diseñado principalmente integrando el módulo de minería de reglas de inducción en la arquitectura del agente inteligente con capacidad de razón en una base de conocimiento de gran escala.

La figura 15, representa en organigrama resultado de la descripción de los aportes mencionados en la anterior categorización.

Figura. 15. Organigrama sección clustering.



Fuente: elaborada por el autor.

3.6. Modelos predictivos

La predicción dentro de las tareas de minería de datos es de las más relevantes, la misma se puede realizar de diversas maneras, en el caso de este trabajo se encontraron producto del análisis de los aportes las categorías de clasificación y regresión. A continuación se relacionan y clasifican en subcategorías propias del análisis los aportes encontrados en la revisión de la literatura que coinciden con estas categorías, iniciado con las relacionadas con tareas de clasificación:

3.6.1. Clasificación.

En el análisis de los aportes encontrados en la revisión de la literatura las tareas de clasificación se pueden categorizar en cuatro ramas que se describirán a continuación así como algunos de los aportes relacionados en cada categoría.

3.6.1.1. Clasificación orientada a la predicción de fraude.

De dos tipos, fraude con tarjeta de crédito y fraude financiero por parte de las empresas. Entre las técnicas convencionales que se aplicaron por parte de los autores en este tipo de tareas se encuentran: regresión logística, máquina de soporte vectorial, bosques aleatorios “Random forest". A continuación se describen algunos aportes en esta categoría.

En referencia al fraude con tarjetas de crédito el trabajo de (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011) se aplican técnicas de minería de datos con regresión logística con el fin de optimizar la predicción de fraude con tarjetas de crédito, las técnicas utilizadas fueron: máquinas de soporte vectorial usando una función kernel, la función de clasificación se puede expresar en términos de productos punto de proyecciones de datos de entrada en un espacio de características de alta dimensión, bosques aleatorios (conocidos como un conjunto de árboles de clasificación o regresión, los bosques aleatorios obtienen variación entre los árboles individuales utilizando dos fuentes de aleatoriedad: en primer lugar, cada árbol se construye con muestras separadas de los datos de entrenamiento; en segundo lugar, solo se considera un subconjunto de atributos de datos seleccionados al azar en cada nodo al construir los árboles individuales), en el modelo de regresión logística se utilizó la variable dependiente de tipo binaria, fraude. Se trabajó con datos reales de transacciones de tarjetas de crédito, con etiquetas basadas en el tipo de transacción, compra minorista, adelanto de efectivo y transferencia.

Por otra parte, en cuanto a predicción de fraude financiero el trabajo de (Ravisankar et al., 2011) se centran en la detección de fraude financiero en diversas compañías que invierten en el mercado de valores Chino. Para este propósito se utilizaron algunas técnicas de minería de datos como: (MLFF) red multicapa de alimentación de red neuronal, (SVM) máquinas de soporte vectorial, programación genética (GP), método de grupo de manejo de datos (GMDH), regresión logística (LR) y red neural probabilística (PNN). El conjunto de datos se trabajó sobre una base de datos real tomada de 202 empresas que cotizan en diversas bolsas Chinas, con 35 ítems financieros relacionados. Cada una de las técnicas se probó en el conjunto de datos, de igual manera se comparan teniendo en cuenta o no la selección de características. De ahí se obtuvieron resultados que muestran que PNN supera a todas las técnicas sin selección de características, mientras que en el caso de técnicas con selección de características las que sobresalieron fueron GP y PNN presentando mayores grados de precisión.

Si bien las técnicas de uso convencional son aplicadas en el desarrollo de las tareas de clasificación en esta categoría, existen también algunas combinaciones y técnicas no convencionales en beneficio de optimizar los procesos y mejorar los resultados obtenidos con técnicas convencionales, algunas de estas técnicas no convencionales son: (MLFF) red multicapa de alimentación de red neuronal, método de grupo de manejo de datos (GMDH) y red neural probabilística (PNN).

3.6.1.2. Clasificación orientada a la predicción de comportamiento de consumidores y tiempo de vida útil de los mismos.

En esta categoría resaltan las tareas de clasificación hacia la predicción del comportamiento de los consumidores en diferentes aspectos como: la estimación del tiempo de vida útil del cliente en una empresa, el estudio de las variables que podrían explicar el comportamiento de los usuarios teniendo como referencia las actividades que realizan, la predicción de la tendencia de personalidad basada en las interacciones entre usuarios, el modelamiento de la respuesta de los

usuarios conociendo el comportamiento de los datos y la minería de relación con los clientes. Los autores aplicaron diversas técnicas de minería de datos hacia el objetivo mencionado en este aparte como las siguientes: regresión logística, árbol de decisión, cadena de Markov, redes neuronales, metodología CHAID, herramientas WEKA. Los sectores hacia donde se orientaron las investigaciones se enmarcan dentro de: Industria de vehículos, centros comerciales, redes y medios sociales, sector bancario y en el desarrollo de campañas de comercialización de productos. A continúan se relacionan y describen brevemente algunos de los aportes:

En referencia a la estimación del tiempo de vida útil de los clientes resalta el trabajo de (Cheng, Chiu, Cheng, & Wu, 2012) en donde se buscaba la predicción del tiempo de vida útil de los consumidores en el sector de mantenimiento de vehículos. Para esto se aplicaron técnicas de minería de datos como: regresión logística y modelo de árbol de decisión (para estimar la probabilidad de abandono de un cliente y predecir la duración de vida útil), una cadena de Markov (para modelar probabilidades de cambio de comportamiento del cliente), redes neuronales (dos redes para predecir el beneficio aportado por el cliente) y el desarrollo de un modelo de predicción LTV (valor del tiempo de vida). Se realizó el estudio sobre base de datos real de la empresa Nissan Motor en Taiwan, incluye datos de ventas y comercialización de vehículos, ventas de autopartes, reparaciones de automóviles y mantenimiento.

Por otra parte, uno de los enfoques encaminados a predecir conductas desarrolladas en centros comerciales por parte de los consumidores se pueden apreciar en el trabajo de (Delgado, Mata, Yepes-Baldó, Montesinos, & Olmos, 2013) se aplica minería de datos al estudio de las variables que explican el comportamiento de los usuarios en centros comerciales, la identificación de los perfiles de los consumidores (teniendo en cuenta las actividades que realizan en los centros comerciales), se utilizó la metodología de segmentación CHAID para la identificación de los perfiles.

El trabajo de (Ortigosa, Carro, & Quiroga, 2014) se centra en la predicción de la tendencia de personalidad basado en las interacciones de usuarios en la red social Facebook, se aplica minería de datos en redes sociales y se desarrolló TP2010, una aplicación en Facebook para recopilar información sobre rasgos de personalidad, mediante técnicas como clasificador Naive Bayes, k-nn, arboles de clasificación y reglas de asociación, se utiliza la herramienta WEKA para la implementación del algoritmo y análisis discriminante a través de la regla lineal de Fisher (análisis gráfico, como resultado el mejor rendimiento se obtuvo al usar todos los atributos recolectados por TP2010, la precisión fue superior al 70% para todos los rasgos, teniendo en cuenta un intervalo de confianza de probabilidad del 99%. La sociabilidad es el rasgo que se predice más exactamente, cerca del 72% de precisión.

Un aporte que resalta por su desarrollo e innovación es el trabajo de (Z.-Y. Chen, Fan, & Sun, 2015) donde se propone un marco de aprendizaje de conjunto jerárquico para modelar la respuesta del usuario en medios sociales, consciente del comportamiento y utilizando diversos datos heterogéneos. Se utiliza la técnica de máquina de soporte vectorial kernel múltiple jerárquica

mejorada (H-MK-SVM) para integrar los datos conductuales, de comportamiento, de compromiso externo, de etiqueta y palabra clave en la selección de características de múltiples atributos correlacionados y para el aprendizaje de conjunto en el modelado de respuesta del usuario.

En el aporte de (Bahari & Elayidom, 2015) se aborda un modelo de minería de relación con los clientes. El modelo se usa para predecir comportamiento de los clientes, mejorar los procesos de toma de decisiones y retener clientes valiosos. El conjunto de datos utilizados contiene los resultados de las campañas de comercialización de entidades bancarias directas, incluye 17 campañas disponibles entre mayo de 2008 y noviembre de 2010, contiene 16 variables de entrada. Se aplican conceptos de minería de datos CRM (Gestión de relaciones con los clientes), redes bayesianas, redes neuronales, WEKA, red neural de percepción multicapa (MLPNN). Aplicados en dos modelos de clasificación para predecir el comportamiento del consumidor. El modelo que logra un alto rendimiento predictivo fue MLPNN con una tasa de precisión del 88,63%.

Como aporte principal dentro de esta categoría se resaltan algunas técnicas no convencionales y combinación de técnicas como: clasificador Naive Bayes, k-nn, arboles de clasificación y reglas de asociación, una modificación de máquina de soporte vectorial kernel múltiple jerárquica mejorada (H-MK-SVM) y una red neural de percepción multicapa (MLPNN), marcan la evolución de las técnicas ya ampliamente utilizadas. Estas técnicas no expuestas en la bibliografía general constituyen los principales avances y establecen la frontera de conocimiento en cuanto a la aplicación de técnicas de minería aplicada a procesos de clasificación.

3.6.1.3. Clasificación orientada a predicción de mercados.

Se aborda en este apartado la predicción de nuevos tipos de mercados, predicción de necesidades del mercado laboral, la clasificación de estrategias de mercadeo exitosas, predicción del éxito de agentes de mercado, segmentación de consumidores en mercadeo, publicidad móvil y la fuga de clientes en los mercados son las principales categorías halladas en la revisión de la bibliografía. A continuación se relacionan algunos aportes:

Un aporte relacionado con nuevos tipos de mercados esta dado en (Warkentin, Sugumaran, & Sainsbury, 2012) donde se realiza la predicción de nuevos tipos de asociaciones de mercados electrónicos mediante el uso de agentes inteligentes, se propone un enfoque de minería de datos basada en agentes genéricos GAMA (arquitectura genérica para agentes basados en minería de datos), el prototipo fue desarrollado usando una interfaz de programación de aplicación (API) en conexión con eBay. Se aplica el estudio a base de datos real, sobre las fuentes de datos de eBay y otras organizaciones.

Un aspecto relevante es la clasificación de estrategias de vendedores es la predicción del éxito de los mismos en diferentes campañas, uno de los aportes relacionados se presenta en (Gordillo-Ruiz, Martínez-Miranda, & Stephens, 2012) en este se busca la clasificación de las estrategias de agentes exitosos en mercados financieros y predecir cuales agentes tendrán mayor éxito. Se aplica análisis bayesiano mediante el método de Bayes ingenuo con el objeto de clasificar las regiones de espacio

multidimensional tomando las estrategias de negocios y ventas de diversos agentes. Se comprobó que las ganancias de los agentes de mercado más exitosos no se deben a suerte sino a estrategias bien elaboradas. La aplicación se efectuó a bases de datos reales de series de tiempo de agentes en mercados financieros, teniendo en cuenta transacciones en un tiempo determinado.

El aporte de (Barrientos, 2013) propone una serie de pasos basados en el proceso de descubrimiento de conocimiento KDD para la correcta integración de los datos, teniendo en cuenta las múltiples plataformas sobre las cuales se accede a la información en el estudio. Se centra en una metodología para la predicción de fuga de clientes en el mercado de las telecomunicaciones, se aplicaron algunas técnicas de minería de datos como: redes neuronales, maquina soporte vectorial (utilizadas para validar y predecir fugados), árbol de decisión (para la clasificación de nuevos planes de oferta a los consumidores), se aplicó segmentación mediante un algoritmo de clúster denominado Two-step clúster. Los datos usados en el proceso son de tipo real, tomados de bases de datos de empresas de telecomunicaciones referentes a: reclamos comerciales, facturación, reclamos técnicos, empresas rivales, detalles de clientes y suscriptores, un total de 208 variables distribuidas en las bases de datos.

Un enfoque de minería de datos para la predicción de las necesidades del mercado laboral se presenta en (Alsultanny, 2013) se aplicaron técnicas como clasificadores Bayesianos, arboles de decisión y reglas de asociación. La técnica Bayes ingenuo (Bayes Naive) se utilizó para crear tablas de entrenamiento, los árboles de decisión se construyeron a partir de un conjunto de datos creado por un proceso conocido como división en el valor de los atributos y las reglas de decisión, con 16 reglas para predecir los mercados de laborales. Este enfoque que pretende identificar y predecir las necesidades de los mercados laborales facilitando un mejor manejo de la información y un adecuado proceso de toma de decisiones. Para las fuentes de información se crearon conjuntos de datos aplicados a empleados que permitió establecer una tabla de conjunto de datos.

Por otra parte, en (K. Y. Kim & Lee, 2014) se buscó la segmentación de consumidores en mercadeo y publicidad móvil, se aplican metodologías y prácticas hibridas así como modelos empíricos Q y R. Los modelos de tipo Q referencia un enfoque típico de investigación cualitativa que busca descubrir e interpretar propiedades internas tales como sentimientos, preferencias, emociones y los modelos de tipo R describe un tipo de estudio empírico o cuantitativo relativamente simple, la variable de R consiste en elementos mensurables o estímulos. Los resultados apoyan los sistemas de relación con los clientes así como los procesos de toma de decisiones.

En los marcos de las observaciones anteriores es manifiesto la aplicación de ciertas técnicas de uso convencional trabajando de manera conjunta, como lo son: clasificadores Bayesianos, arboles de decisión, redes neuronales, maquina soporte vectorial, se presentan también el uso de metodologías ampliamente abarcadas como es el caso de los procesos de KDD y otras que resultan novedosas como las relacionadas con la minería basada en agentes genéricos y las metodologías hibridas.

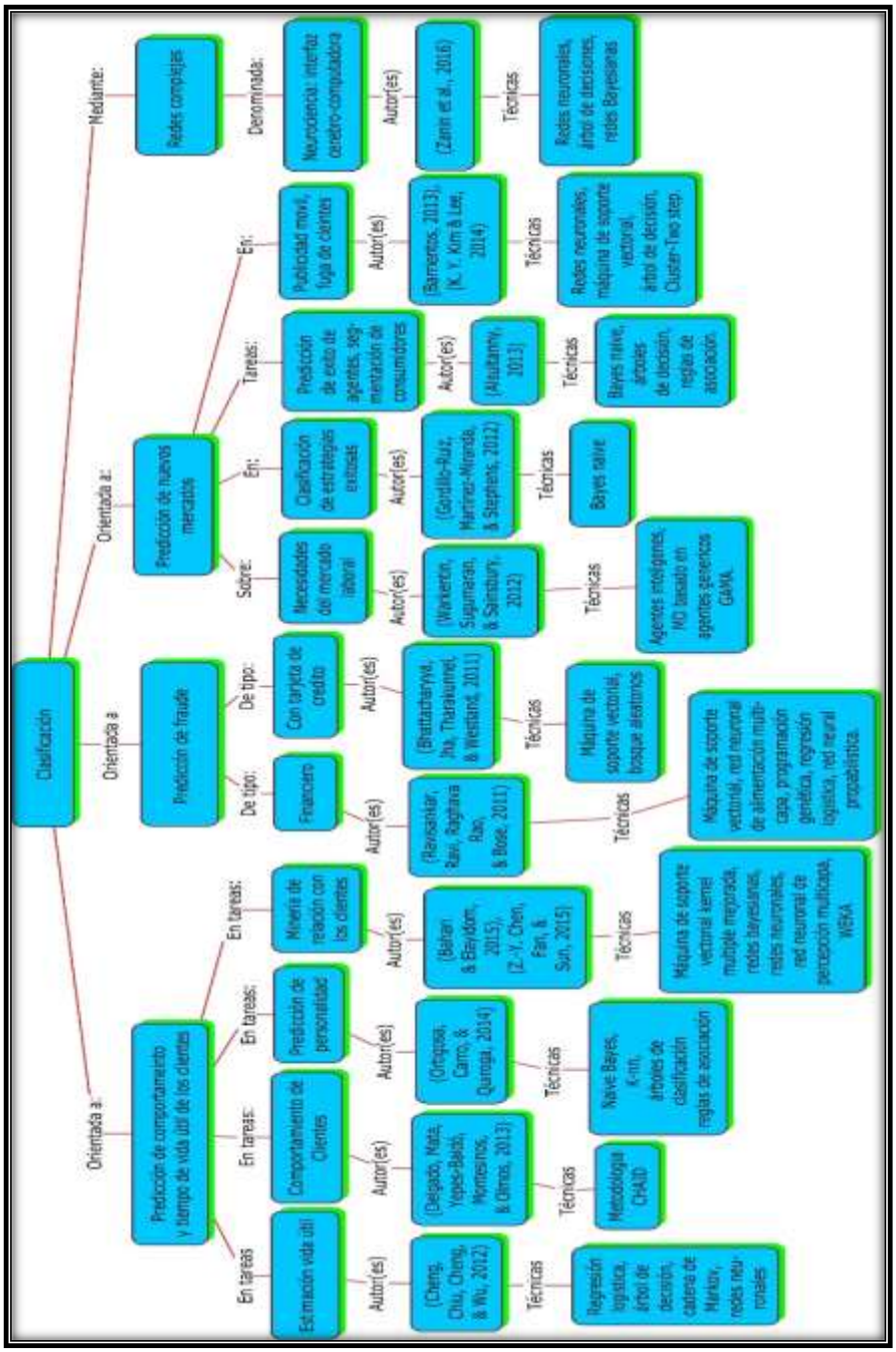
3.6.1.4. Clasificación con redes complejas.

En esta categoría, se aplicó la teoría de redes complejas en conjunto con las técnicas de extracción de datos relacionadas con la minería de datos, se aplicaron técnicas propias de minería como redes neuronales artificiales, árbol de decisión, redes bayesianas. A continuación se describe el trabajo de (Zanin et al., 2016).

El rumbo del trabajo se centró en la clasificación con redes complejas denominada neurociencia con interfaz cerebro computadoras (BCI). El desarrollo se fundamentó en la combinación de redes complejas y extracción de datos mediante técnicas de minería, se aplicaron técnicas de análisis de redes complejas, redes neuronales artificiales, árbol de decisión y redes bayesianas en análisis omicos, referido a campos de la biología que terminan en omicos, ejemplo estudios genómicos. Los resultados del estudio son aplicables a cualquier área de negocios y como apoyo a los procesos de toma de decisiones.

En la figura 16, se puede apreciar el organigrama resultado de la sección relacionada con técnicas de clasificación y la categorización descrita anteriormente.

Figura. 16. Organigrama sección clasificación.



Fuente: elaborada por el autor.

3.6.2. Regresión

Las técnicas de regresión son ampliamente utilizadas en procesos de predicción y clasificación, entre estas: las redes neuronales, regresión lineal, arboles de decisión y otras, siempre apoyadas en el análisis estadístico como pilar fundamental de desarrollo, la regresión suele relacionarse con pruebas de software que intentan descubrir diferentes aspectos como causas de errores, deficiencias en la funcionalidad y cambios realizados en los procesos o aplicaciones. A continuación se relacionan algunos de los aportes encontrados en la revisión de la literatura.

3.6.2.1. Regresión orientada hacia predicción de puntajes de influencia de revisores.

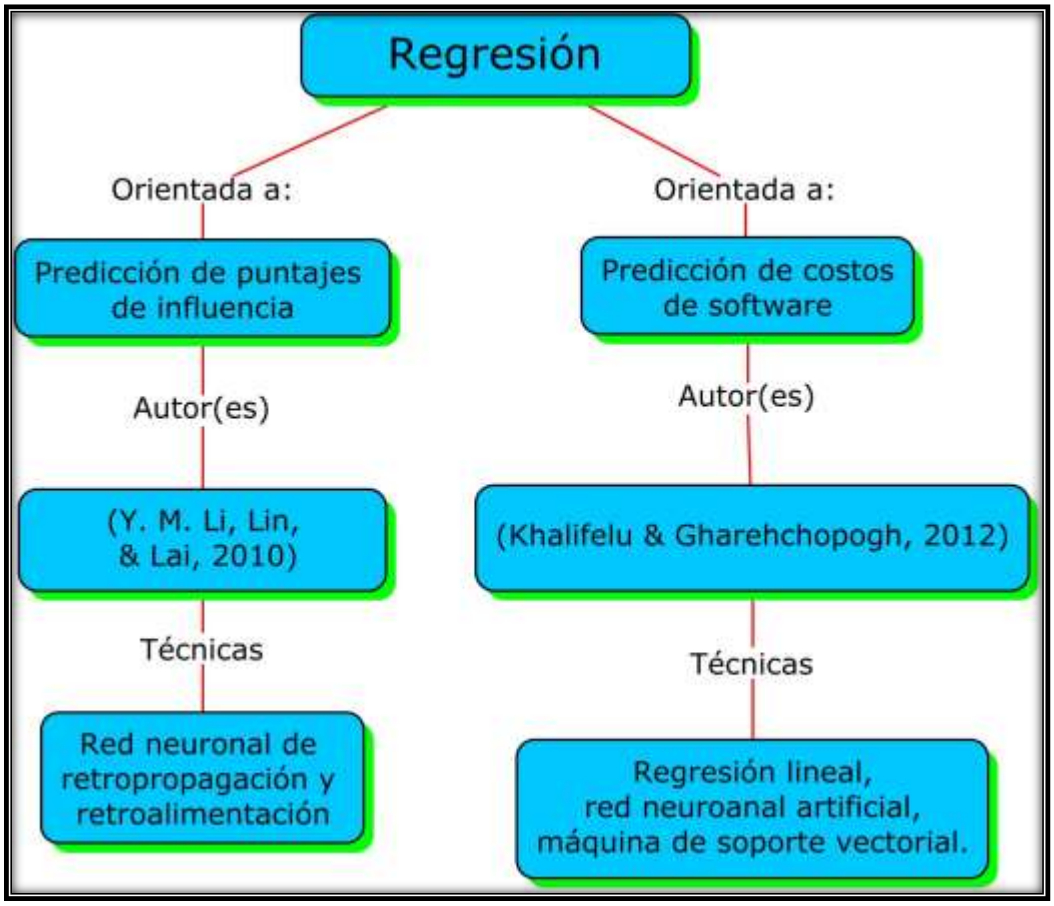
Un trabajo relacionado con uno de los aspectos claves en el mercadeo es la manera como se logre difundir los productos, en esta área, el aporte de (Li, Lin, & Lai, 2010) se centra en predecir el puntaje de influencia de revisores para el sector de mercadeo en línea de boca a boca, para lograr este objetivo se aplica una red neuronal de 3 capas de retro propagación y retroalimentación. La capa 1 contiene 3 neuronas, la capa oculta es el componente de procesamiento central generando la estructura de la red, la capa 3 contiene una neurona la cual es la encargada de mostrar el resultado. Los datos son tomados del servidor web Epinions.com, una plataforma abierta que ofrece reseñas generadas por los usuarios sobre diversos tipos de productos, los resultados pueden orientarse hacia el desarrollo de estrategias de ventas.

3.6.2.2. Regresión orientada a la predicción de costo de software.

La estimación adecuada del costo de producción de software es una de las fases primordiales a la hora de realizar cualquier tipo de desarrollo, trabajos orientados en este tema se adelantó en (Khalifelu & Gharehchopogh, 2012) se buscaba mejorar la predicción de estimación de costo usando métodos de minería de datos, para esto se realiza la comparación y evaluación de técnicas de minería en modelos de estimación de costo de software como el Modelo de Costo Constructivo (COCOMO), se utilizan datos sintéticos para las pruebas de conjuntos de datos históricos tomados de proyectos anteriores del Instituto de Tecnología de California y el laboratorio de propulsión a chorro de la NASA, se aplican técnicas de regresión lineal, red neuronal artificial, máquina de soporte vectorial, k-vecino cercano (usados en la categorización del clasificador de minería de datos para el modelo de estimación de costos) y WEKA. Los resultados del trabajo orientan y sirven de apoyo a los procesos de planificación y desarrollo de software.

En la figura 17, se presenta el organigrama de la sección regresión, resultado de la descripción y categorización de los aportes encontrados en la literatura.

Figura. 17. Organigrama sección regresión.



Fuente: elaborada por el autor.

3.7. Reglas de asociación

Esta categoría involucra los aportes relacionados con la aplicación de reglas de asociación en diversos sectores y enfoques así como el trabajo conjunto con otras técnicas. Se pueden establecer cuatro ramas producto del análisis de los aportes recopilados en la bibliografía consultada: la primera rama, orientada hacia la aplicación de reglas de asociación en tareas de clasificación, dentro de esta, clasificación de precios de productos y clasificación de preferencias y comportamientos de los usuarios. La segunda rama, comprende las tareas de predicción con base en la aplicación de reglas de asociación, encaminadas hacia predicción de eventos meteorológicos, sugerencias y soluciones para empresas de compra en línea y portafolios de inversión en mercados de valores, se presenta también una mejora en la predicción de las preferencias de los usuarios y la predicción de órdenes de prescripción en farmacias. La tercera rama, se encamina hacia el descubrimiento de patrones ocultos y patrones frecuentes. La última rama, involucra otros aspectos como lo son: la extracción de reglas coherentes difusas y la identificación de ponderaciones casuales. A continuación se describirán algunos de los aportes relacionados en cada una de las ramas mencionadas.

3.7.1. Reglas de asociación orientadas a la clasificación.

Un caso típico donde las tareas de clasificación son relevantes es en la adecuada estimación de los precios de los productos en estanterías, la misma se debe realizar teniendo en cuenta el cambio dinámico que estos presentan, para esto en (Nafari & Shahrabi, 2010) se desarrolló un nuevo enfoque en donde se aplican reglas de asociación multinivel para encontrar las relaciones entre los productos con respecto a sus precios, estas se aplican en conjunto con una modificación del algoritmo RApriori-TdMl. Para el desarrollo del trabajo se usaron datos reales, tomados de la venta al por menor de una cadena de supermercados en Irán, donde cada transacción registra las siguientes variables: tiempo de compra, artículos comprados, los precios, descuentos y cantidades. Ente enfoque ofrece apoyo a los procesos de desarrollo de estrategias de ventas y toma de decisiones.

Dentro de la misma rama otro enfoque se presenta en (Wen, Liao, Chang, & Hsu, 2012) este trabajo se orienta a la clasificación de las preferencias y comportamiento de los usuarios en el mercado de productos de lujo. Se analiza teniendo como base la minería de comportamiento de compras, se usa un enfoque de reglas de asociación que comprende el algoritmo apriori, reglas y mapas de conocimiento junto con el algoritmo k-means. Los datos son recolectados mediante cuestionario, constituyendo una base de datos reales que contenía siete partes: información del cliente, estilo de vida individual, información del canal de ganancia, experiencia individual de compra de productos de lujo, actividades de promoción de productos, cognición de marca y la percepción del consumidor. Los resultados del análisis se presentan como mapas de conocimiento que permiten a los gerentes de las compañías prever y formalizar campañas de ventas.

Resaltan en esta rama la aplicación de reglas de asociación multinivel, reglas y mapas de conocimiento en conjunto con técnicas de clustering como k-means y la mejora en los métodos de asociación tradicional mediante enfoques de clasificación de tipo asociativa, un enfoque de minería de asociación principal, sin dejar de lado el uso de algoritmos convencionales como el Apriori.

3.7.2. Reglas de asociación orientadas a la predicción.

En relación con la predicción de eventos meteorológicos se presenta el trabajo de (Peng et al., 2011), el objetivo es prever mediante un marco de gestión de incidentes las emergencias y eventos futuros basados en información de incidentes presentados anteriormente y que se encuentren disponibles para el análisis. Se aplican técnicas de minería de datos como: minería de patrones frecuentes, reglas de asociación y análisis de clúster. Para el análisis se usan datos reales, tomados del registro de 5 años de desastres agro meteorológicos que ocurrieron en China desde 1997 hasta 2001, con cinco bases de datos de la misma estructura, cada una reúne ocho atributos para representar el daño a los cultivos en 31 provincias chinas causadas por cuatro tipos de desastre agro meteorológicos. Los resultados sirven como apoyo en el proceso de toma de decisiones de los cultivadores en esta región.

De acuerdo con los razonamientos que se han venido realizando en esta misma rama el objetivo del trabajo de (Liao, Chu, Chen, & Chang, 2012) se centra en la predicción de sugerencias y soluciones para empresas de compra grupal en línea. Para el desarrollo se propuso un enfoque de minería de datos que permite explorar el comportamiento de compra, se emplea el análisis bajo un algoritmo a priori con enfoque de reglas de asociación y clustering. Utilizó un conjunto de datos reales y se maneja el enfoque de cuestionario para recopilar datos de clientes. El cuestionario está dividido en cinco partes; la parte uno se centra en la información personal del cliente, la parte dos discute la psicología de los clientes individuales que participan en la compra, la parte tres las tendencias del cliente individual de la compra en grupo, mientras que la parte cuatro se centra en el comportamiento del individuo y la parte final el mecanismo de servicio, resultando de este proceso dos clúster; un grupo de consumidores potenciales y otro grupo con consumidores objetivo, de esta manera se obtiene conocimiento del cliente dentro de un grupo de clientes compradores. En total, 720 cuestionarios fueron enviados y 621 fueron recogidos, excluyendo omisiones y respuestas incompletas. Los resultados permiten mejorar las estrategias de retención de clientes valiosos identificados en los grupos de compra en línea.

En el trabajo de (Liao & Chou, 2013) se analiza la predicción de posibles portafolios de inversión en mercados de valores, en este sentido se construye un sistema de base de datos para almacenar y consultar efectivamente sobre un gran número de transacciones; La base de datos se analiza mediante un esquema de estrella que consta de dos tipos de tablas: tablas de hechos y tablas de dimensiones, las tablas de hechos contienen datos fácticos o cuantitativos y las tablas de dimensiones tienen datos descriptivos. Se aplicaron herramientas y técnicas como: SPSS modeler, reglas de asociación y análisis de clúster, se analizaron 30 categorías de índices bursátiles y se implementan como variables de decisión para observar el comportamiento de las asociaciones de relevancia en mercados de valores, incluye índices de recopilación de categorías de índices industriales, la fuente de datos fue el sitio web de la revista de economía de Taiwán.

Con el ánimo predecir mejor las preferencias de los clientes en el sector de mercadeo móvil resulta un concepto nuevo para la captura de información contextual apoyado en un método de post-poda que permite eliminar las reglas redundantes, es un enfoque genérico para emparejamiento no consecutivo de los umbrales de tiempo mediante reglas secuenciales multidimensionales, en concordancia se presenta en (Tang, Liao, & Sun, 2013) un nuevo marco con un procedimiento de tres etapas para descubrir la correlación entre los contextos de los usuarios de dispositivos móviles y sus actividades. Este enfoque basado en minería de datos se utiliza para extraer y aplicar reglas secuenciales como una solución de mercadeo móvil personalizado (MPM), así mismo habilita la incorporación de información contextual multidimensional al interior de minería de reglas secuenciales. Los datos utilizados fueron tomados de la base de datos de contexto denominada "Datos de contexto de Nokia". El conjunto de datos consiste en una secuencia de datos contextuales de 43 diferentes registros de sesión.

Un enfoque encaminado hacia la predicción de órdenes de prescripción específicamente en las farmacias de surtido central que distribuyen medicamentos a farmacias minoristas enmarcadas en

el área de farmacias automatizadas se presenta en (Khader, Lashier, & Yoon, 2016). Se utilizan técnicas de minería de reglas de asociación y el algoritmo FP-growth de patrones de crecimiento frecuente para extraer los conjuntos de artículos frecuentes de medicamentos. El objetivo fundamental de la investigación fue mejorar las estrategias en la automatización de farmacias mediante el uso de un enfoque de minería de datos. Los datos fueron tomados de una base de datos real de recetas de farmacia, en una gran farmacia central de los Estados Unidos.

Las reglas de asociación representan un gran aporte en cuanto a las tareas de predicción teniendo en cuenta los aportes en esta rama se evidencia la aplicación conjunta de técnicas como: minería de patrones frecuentes, reglas de asociación y análisis de clúster apoyando la solución de problemas específicos, también la aplicación de herramientas como el SPSS modeler en trabajos conjuntos con técnicas de clustering principalmente. El desarrollo de nuevos enfoques que permiten mejora en cuanto a la eliminación de reglas redundantes y la extracción de reglas secuenciales resaltan por los esfuerzos en la evolución de las mismas técnicas, así como trabajos conjuntos entre reglas de asociación y algoritmos de patrones de crecimiento frecuente.

3.7.3. Reglas de asociación orientadas al descubrimiento de patrones.

El descubrimiento de patrones ocultos en el sector de mercadeo y ventas es de las principales tareas en los sistemas de inteligencia de negocios, el trabajo que se presenta en (Cheung & Li, 2012) aplica un método de extracción cualitativa de coeficientes de correlación que es capaz de descubrir patrones ocultos de ventas y mercadeo, para alcanzar este objetivo se implementó un prototipo de sistemas de negocios inteligente en conjunto con un algoritmo de minería de datos de ventas de coeficientes de correlación (CCSDMS) que se aplicó con éxito en un sitio de referencia especificado. Los datos usados en el trabajo fueron tomados de una base de datos real específicamente de un sistema de gestión de información (MIS) que incluye información comercial, perfiles de clientes, información de productos, presupuestos de clientes y ventas, los datos se almacenan en formatos fijos y pueden ser exportados a través de SQL. Los resultados de aplicación del método propuesto evidencian mejor poder predictivo, mayor precisión y más eficiencia computacional.

Para dar continuidad en la rama de aportes que relacionan reglas de asociación y clasificación en el trabajo de (F. Chen, Wang, Li, Wu, & Tian, 2014) se presenta un nuevo enfoque de clasificación asociativo eficiente que busca mejorar los problemas de métodos de asociación tradicionales en cuanto a la producción de reglas de clases redundantes, se basa en una métrica de calidad de regla denominada principado que mide la precisión y cobertura de una regla para una clase especifica. Se aplica el enfoque de la minería de asociación principal (PAM) que consta de cuatro fases: descubrimiento de patrones frecuentes, generación de reglas, poda de reglas y clasificación de objetos, en conjunto con la herramienta WEKA. La clasificación asociativa produce un clasificador compacto con un pequeño número de reglas de asociación. Los datos para el análisis y experimentos son tomados de 17 conjuntos de datos del repositorio de la máquina de aprendizaje.

3.7.4. Reglas de asociación orientadas a la extracción de reglas e identificación de relaciones casuales.

Un aporte relevante en cuanto a la extracción de reglas difusas se presenta en el trabajo desarrollado por (C. H. Chen, Li, & Lee, 2013), en este se propone un algoritmo para extraer reglas coherentes difusas y superar problemas con las propiedades de la lógica proposicional. El algoritmo primero transforma las transacciones en conjuntos difusos, luego se recopilan los conjuntos para generar reglas coherentes difusas, posteriormente se calculan tablas de contingencia para verificar si las reglas candidatas satisfacen ciertos criterios, si los cumplen, se considera la regla como reglas coherente difusa. Se usaron los datos de la base de datos de foodmart que se encuentra en el producto de base de datos Microsoft SQL Server 2000. Los resultados experimentales mostraron la eficiencia del algoritmo propuesto sobre la base de datos foodmart.

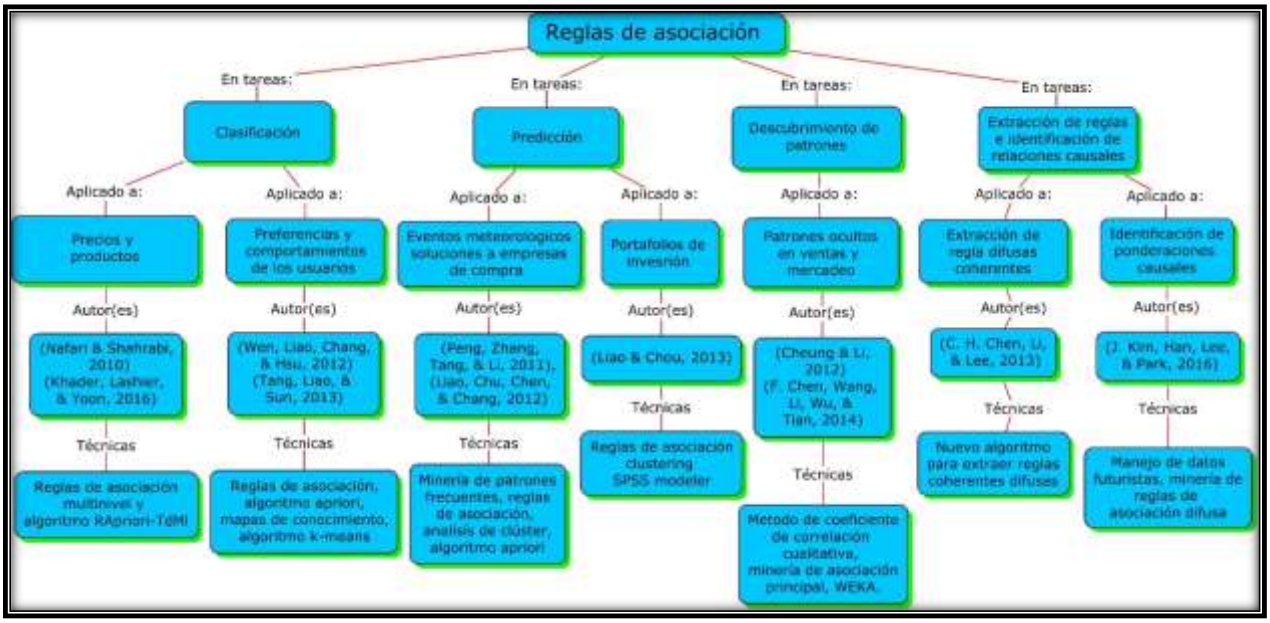
En correspondencia con la identificación de relaciones casuales y con el objetivo de extraer conceptos de escenarios de datos futuristas en documentos textuales mediante la integración de técnicas en escenarios de manejo de este tipo de datos que constituye un novedoso campo de desarrollo, el aporte de (J. Kim et al., 2016) se desarrolló sobre la utilización de datos futuristas, una colección de opiniones orientadas al futuro extraídas de las comunidades en línea. Se aplicó la técnica de minería de reglas de asociación difusa (FARM) la cual se utiliza para identificar ponderaciones causales basadas en las reglas si-entonces en conjunto con la integración de mapas cognitivos difusos (FCMs) los cuales se constituyen una técnica para el desarrollo de escenarios y conceptos relacionados con el futuro, por otra parte, minería de texto y el análisis semántico latente (LSA) aplicados con el objeto de extraer conceptos en documentos textuales.

Como aporte final a esta rama se presenta el trabajo de (Sahoo, Das, & Goswami, 2015) en el cual se propuso una representación comprimida para las reglas de asociación que tengan el mínimo antecedente y el consecuente máximo, la cual se genera con ayuda de conjuntos de elementos de alta utilidad. Se proponen algoritmos para generar reglas y métodos de asociación no redundantes como el algoritmo HUCI-Miner que identifica los conjuntos de elementos cerrados de alta utilidad y sus generadores y permite la generación eficiente de las reglas de asociación no redundantes entre los conjuntos que utilizan el marco de confianza de la utilidad.

Tomando como referencia las contribuciones anteriores de los diferentes autores resaltan algunas aplicaciones donde se utilizan algoritmos como los que permiten la extracción y el manejo de reglas coherentes difusas, además de los aportes que contribuyen con nuevos campos de desarrollo como es el caso del manejo de datos y escenarios futuristas y algoritmos que generan reglas y métodos de asociación no redundantes como el HUCI-Miner.

En la figura 18, se puede apreciar el organigrama de la sección como resultado de la clasificación desarrollada.

Figura. 18. Organigrama reglas de asociación.

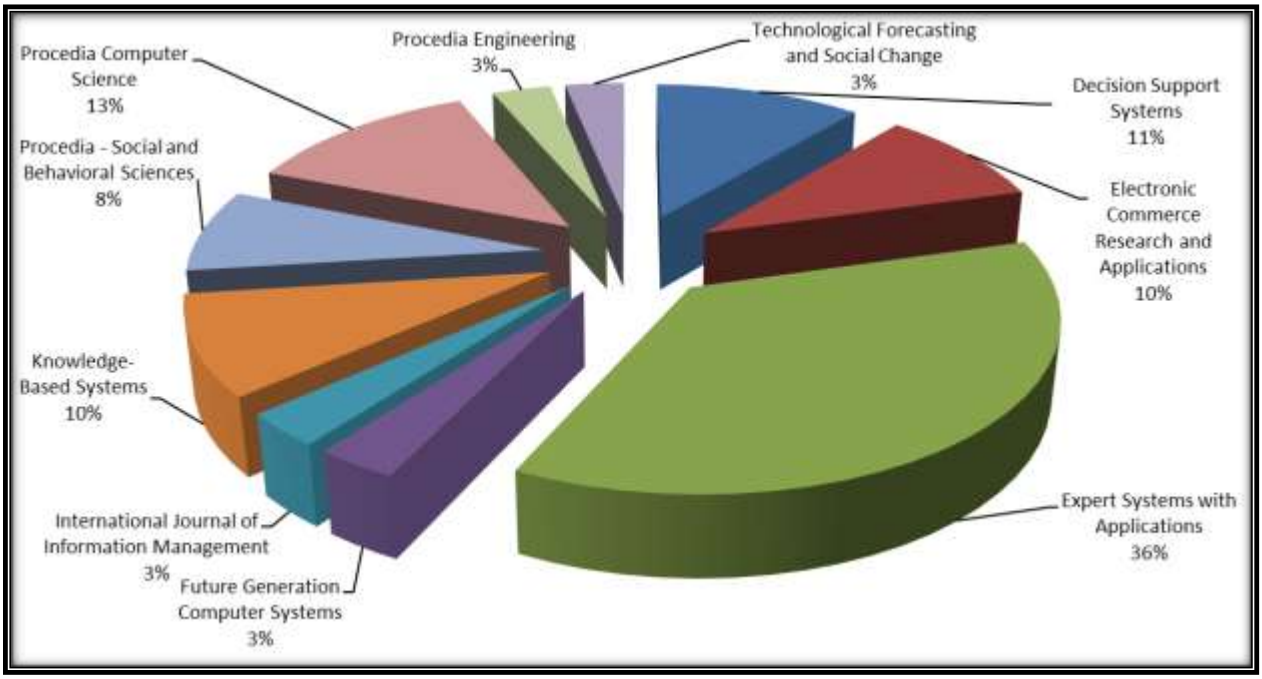


Fuente: elaborada por el autor.

3.8. Revistas con mayores aportes

Teniendo en cuenta la figura 19, se relacionan las revistas en donde los autores divulgaron sus trabajos, el mayor porcentaje de aportes de los autores está en la revista: “Expert Systems with Applications” con 36%, la siguiente es “Procedia Computer Science” con un porcentaje de participación del 13%, continua “Decision Support Systems” correspondientes al 11%. La siguen las revistas “Knowledge-Based Systems” y “Electronic Commerce Research and Applications” cada una con un 10 % de los aportes, a continuación esta la revista “Procedia - Social and Behavioral Sciences”, corresponde a un 8%, para finalizar con un porcentaje del 3%, están las revistas: “Technological Forecasting and Social Change”, “International Journal of Information Management”, “Procedia Engineering”, y “Future Generation Computer Systems”. Es relevante el amplio espectro de revistas en las que los autores han podido realizar la divulgación de sus aportes, esto constata que estas áreas de conocimiento son de gran interés por parte de la comunidad científica.

Figura. 19. Revistas con mayores aportes.

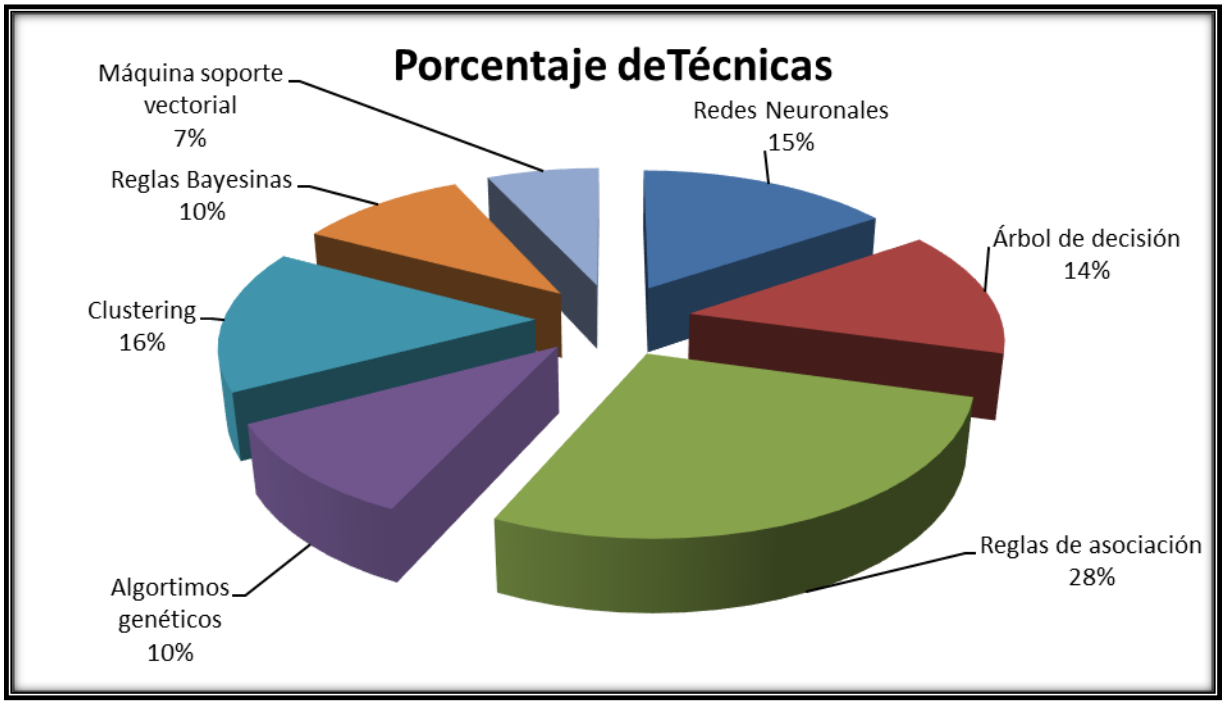


Fuente: elaborada por el autor.

3.9. Técnicas convencionales más utilizadas

Como se puede apreciar en la figura 20, la técnica con más aceptación y apropiación son las reglas de asociación con un 28%, los siguen las técnicas de clustering con un 16%, muy de cerca se encuentran las redes neuronales y las técnicas de árbol de decisión con un 15% y 14% respectivamente, las reglas Bayesianas y los algoritmos genéticos ocupan un 10% y finaliza las técnicas de máquina de soporte vectorial con un 7%, de esta manera se establece el orden de aplicación y apropiación de las técnicas de minería de datos más utilizadas en pro del apoyo a las soluciones de inteligencia de negocios.

Figura. 20. Porcentaje de técnicas.



Fuente: elaborada por el autor.

Unos de los aspectos más relevantes dentro del desarrollo del estado del arte de la minería de datos aplicada a inteligencia de negocios se puede apreciar en el cruce de las áreas de inteligencia de negocios y las técnicas de minería de datos, como se puede apreciar en la tabla 2.

Tabla 2. Cruce técnicas minería de datos, áreas de inteligencia de negocios.

Técnicas minería de datos	Áreas de inteligencia de negocios				
	Clientes	Proveedores	Productos	Servicios	Competidores
CLUSTERING	27.7%	20%	35.3%	17.4%	20.8%
CLASIFICACIÓN	38.8%	20%	35.3%	39.1%	45.8%
REGRESIÓN	11.1%	0%	0%	8.7%	16.65
REGLAS DE ASOCIACIÓN	22.2%	60%	29.4%	34.7%	16.6%
TOTAL	100%	100%	100%	100%	100%

Fuente: elaborada por el autor.

Como se puede observar en la matriz de la tabla número 2, se presentan las áreas de inteligencia de negocios tomadas como criterios para selección y análisis de los artículos y las técnicas de minería de datos tomadas en cuenta en el análisis y clasificación, es posible identificar el porcentaje de aportes en cada una de las áreas y el cruce con las categorías establecidas dentro de las técnicas de minería de datos, en el área relacionada con competidores la mayor cantidad de aportes se centran en las técnicas de clasificación con un 45.8% de los aportes; en el área de servicios las

técnicas de clasificación y reglas de asociación son las predominantes con 39.1% y 34.7% de los aportes; con respecto al área que hace referencia a clientes predominan las técnicas de clasificación con un 38.8% de los aportes; en el área de productos los aportes se encuentran distribuidos más uniformemente entre las técnicas de clustering, clasificación y reglas de asociación, con porcentajes de 35%, 3.5%, 29% respectivamente. Por último el área de proveedores en donde predominan las técnicas de reglas de asociación con un porcentaje de 60%.

Capítulo 4. Conclusiones, resultados, recomendaciones y trabajos futuros

Las conclusiones y recomendaciones del presente trabajo están orientadas hacia la perspectiva de los trabajos futuros en las categorías de clasificación planteadas, se presentan a continuación las líneas de investigación futura en cada una de las categorías seleccionada para el desarrollo del estado del arte de la minería de datos en inteligencia de negocios.

4.1. Conclusiones

En relación con la información recopilada se logró seleccionar y analizar una cantidad de artículos significativa, alrededor de 90 artículos.

Con sustento en el marco teórico establecido fue posible realizar la categorización de los artículos en una clasificación propia del análisis de la información recopilada.

Como producto de la clasificación establecida se pudieron identificar aportes relevantes de las técnicas, herramientas y metodologías de la minería de datos aplicada a los procesos y áreas de inteligencia de negocios.

4.1.1. Técnicas convencionales más utilizadas.

Las técnicas de minería de datos convencionales más consolidadas se enmarcan en las técnicas de clustering, algoritmos K-means, algoritmos de clúster jerárquico, así mismo las técnicas de máquina de soporte vectorial, regresión logística, minería de patrones frecuentes, minería de opinión, minería de texto, minería de reglas de asociación (ASM), método de Bayes ingenuo, K-NN, arboles de clasificación, arboles de decisión, algoritmo de clúster Two-step clúster, redes neuronales, modelo de regresión de clase latente LCRM, modelo de regresión probabilística (PRM), regresión lineal, herramienta WEKA, SPSS modeler, algoritmo RApriori-TdMl, algoritmo FP-growth.

4.1.2. Modificaciones de técnicas tradicionales y técnicas de vanguardia

Como producto del análisis de los aportes recopilados se hizo posible identificar algunas técnicas de vanguardia y algunas modificaciones y mejoras a técnicas tradicionales como lo son: red neuronal de alimentación multicapa, programación genética, método de grupo de manejo de datos, red neuronal probabilística, algoritmos de propagación de etiqueta, algoritmo de mineralización de crecimiento de patrón frecuente e incierto, reglas de asociación multinivel, lenguaje natural de procesamiento, modelo de Markov oculto lexicalizado (L-HMM), modelo de campos aleatorios condicionales (CRF), ASM más reglas lingüísticas basadas en L-HMM en CRF, minería de reglas de asociación difusa (FARM), método de análisis semántico latente (LSA), simulador MACOM (multi-agente basado en mapas cognitivos difusos), mapas cognoscitivos difusos (FCM), métodos de extracción de frases (KEA) y de conceptos claves (ACE), así como una mejora a este último, llamado (ICE), modificación de máquina de soporte vectorial kernel múltiple jerárquica mejorada (H-MK-SVM), red neural de percepción multicapa (MLPNN), minería basada en agentes

genéricos, algoritmo de clúster de líderes, clúster suave, clúster de jerarquía acumulativa, algoritmos K-NN-IR y K-means-IR (modificados para el tratamiento de reglas de inducción), minería de agentes inteligentes (MIA), algoritmo de minería de datos de ventas de coeficientes de correlación (CCSDMS), minería de asociación principal (PAM), algoritmo HUCI-Miner.

Como resultado de realizar cruce entre las técnicas de minería, encasilladas en los modelos descriptivos, predictivos y su relación con las áreas que incluye la inteligencia de negocios se presenta en la tabla numero 2 la matriz correspondiente, en esta se relacionan las técnicas de clustering, clasificación, regresión, reglas de asociación y las publicaciones en relación con la minería de datos e inteligencia de negocios que cumplen con la afinidad de áreas como productos, clientes, servicios, competidores y proveedores.

4.2. Resultados.

Como resultado del desarrollo del trabajo se lograron los siguientes productos:

Artículo: "EL PAPEL DE LA MINERÍA DE DATOS EN LA INTELIGENCIA DE NEGOCIOS, UNAREVISIÓNLITERARIA.http://ciinatic2017.ufps.edu.co/wordpress/MemoriasCIINATIC2017.pdf"

Ponencia: El papel de la minería de datos en la inteligencia de negocios, una revisión literaria. **CONGRESO INTERNACIONAL EN INNOVACIÓN Y APROPIACIÓN DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES. 2017.**

4.3. Recomendaciones

En cuanto a recomendaciones para mejorar este trabajo pueden surgir múltiples desde la perspectiva que se desee abordar el tema, por ejemplo sería recomendable seleccionar una de las áreas que relaciona la inteligencia de negocios y aplicar una revisión centrado en una de estas áreas.

También es recomendable replicar algunas de los trabajos referenciados en un ámbito local o regional con el objetivo de comparar la aplicación de técnicas y métodos en una problemática real y particular.

4.4. Trabajos futuros

4.4.1. Trabajos futuros en clasificación

Futuros desarrollos se pueden orientar sobre atributos más amplios para el modelamiento de fraude con tarjeta de crédito, centrarse en la diferencia de la secuencia de transacciones fraudulentas y legítimas antes de que se retire una tarjeta de crédito, examinar las diferencias en el comportamiento fraudulento entre los diferentes tipos de fraude, por ejemplo, la diferencia de comportamiento entre las tarjetas robadas y falsificadas.

Futuras investigaciones en técnicas de simulación, como una dinámica del sistema o una red de Petri, para modelar el comportamiento de compra futura de un cliente.

Analizar otras fuentes de comportamiento del usuario, como los patrones de movimiento del mouse, que han demostrado estar relacionados con algunas características del usuario, los estilos de aprendizaje, es posible que la extracción de varios conjuntos de información heterogénea conduzca a clasificaciones más precisas de la personalidad del usuario.

Optimizar la evaluación de métricas de red, infraestructuras de software para análisis de redes complejas.

Trabajos próximos pueden centrarse en la aplicación de clasificadores neuro-fuzzy para modelos de gestión de relaciones con los clientes.

4.4.2. Trabajos futuros en clustering

Un trabajo futuro referente al uso de técnicas de minería gráfica y la generación de grupos de productos enmarcado en la personalización de la caracterización del cliente en diferentes grupos, basado en el grado de membresía de cada comunidad a partir de las compras previas de un cliente.

Sistemas inteligentes que permitan agregar más funcionalidad para los servicios de minería web, con el objetivo de hacer que el uso de la web sea más útil en comercio electrónico

Mejorar los enfoques de los algoritmos de recomendación y comparar la recomendación mejorada con otros métodos existentes de recomendación, como el método de filtrado colaborativo (FC).

Implementar nuevas plataformas para minería de agentes inteligentes MIA.

Una dirección futura de investigación puede orientarse hacia el mejoramiento de los procesos de selección de algoritmos de minería de datos, mediante técnicas de selección y clasificación mejoradas, en busca de brindar más información a las empresas para implementar soluciones de inteligencia de negocios, lo anterior y el coste computacional en algunos casos son los factores más críticos para la puesta en marcha de estos procesos.

4.4.3. Trabajos futuros en regresión

Trabajos futuros orientados al desarrollo de mecanismos de ponderación flexible para ajustar los modelos de mercadeo en línea a las necesidades de las empresas y reducir costos.

Aplicar en el futuro algoritmos genéticos, árboles de decisión difusos a proyectos de ingeniería y estimación de costos enfocados en técnicas de regresión lineal múltiple.

4.4.4. Trabajos futuros en reglas de asociación

Trabajos futuros se deben centrar en deducir asociaciones entre diferentes clientes mediante el análisis de patrones de consumo para mejorar las estrategias de venta.

En futuras investigaciones aplicar el algoritmo de minería de reglas coherentes difuso a problemas más complejos para probar su rendimiento.

Futuras investigaciones se centran en filtrar reglas basadas solo en la restricción de confianza.

Un futuro campo de investigación es estudiar cómo definir el principado para una combinación y clasificación de múltiples reglas, la evaluación de la calidad de la regla, la mejora de la estrategia de poda para reducir el tamaño del modelo de clasificación, así como los métodos para la predicción de etiquetas múltiples.

Un sector en el que se debe realizar desarrollos y enfatizar especial cuidado es el concerniente con la privacidad y seguridad de la información de los usuarios, sobre todo cuando estos son insumo para la ejecución de los proyectos de BI, la vista profunda de los marcos legales establecidos y la sinergia con la aplicación de estos sistemas contribuyen al fortalecimiento de los procesos de inteligencia de negocios.

Líneas de investigación próxima están planteadas hacia sistemas de diseño de investigación científica DSR, en aspectos relevantes como la recolección y mezcla de datos, con metodologías de modelado de meta-datos multidimensionales, metodologías para el desarrollo de herramientas o componentes de herramientas en la nube.

Una orientación futura hacia un desarrollo y evolución de un concepto nuevo relacionado con minería de deportes, se orienta a analizar datos históricos de equipos para predicción de resultados en futuras temporadas de juegos.

Una mirada hacia los algoritmos de aprendizaje automático como es caso de Deep Learning y su aplicación en soluciones de inteligencia de negocios es una de las cuestiones de investigación que surgen como proceso de evolución de minería de datos.

Posterior al análisis de la información fue posible identificar tres áreas fundamentales sobre las cuales la minería de datos y la inteligencia de negocios brindan apoyo a las empresas, estas áreas son: toma de decisiones, apoyo al desarrollo de estrategias de ventas y apoyo en la planificación empresarial.

Referencias bibliográficas

Alsultanny, Y. A. (2013). Labor market forecasting by using data mining. In *Procedia Computer Science* (Vol. 18, pp. 1700–1709). https://doi.org/10.1016/j.procs.2013.05.338

Aluja, t. (2001). la mine ia de datos, entre la esta istica y la inteligencia artificial, *25*(3), 479–498.

Amarouche, K., Benbrahim, H., & Kassou, I. (2015). Product Opinion Mining for Competitive Intelligence. *Procedia Computer Science*, *73*(Awict), 358–365. https://doi.org/10.1016/j.procs.2015.12.004

Bahari, T. F., & Elayidom, M. S. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science*, *46*, 725–731. https://doi.org/10.1016/j.procs.2015.02.136

Barrientos, F. S. R. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. *Revista Ingeniería de Sistemas*, *XXVII*, 73–108.

Beltrán Martínez, M. B. (2003). Minería de datos, 67.

Ben-David, S., & Shalev-Shwartz, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. *Understanding Machine Learning: From Theory to Algorithms*. https://doi.org/10.1017/CBO9781107298019

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Cano, J. L. (2007). *Business Intelligence: Competir Con Información*. *Banesto, Fundación cultural*. Retrieved from http://itemsweb.esade.edu/biblioteca/archivo/Business_Intelligence_competir_con_informacion.pdf

Carmona Suárez, E. J. (2014). Tutorial sobre Máquinas de Vectores Soporte (SVM), 1–25.

Chemchem, A., & Drias, H. (2015). From data mining to knowledge mining: Application to intelligent agents. *Expert Systems with Applications*, *42*(3), 1436–1445. https://doi.org/10.1016/j.eswa.2014.08.024

Chen, C. H., Li, A. F., & Lee, Y. C. (2013). A fuzzy coherent rule mining algorithm. *Applied Soft Computing Journal*, *13*(7), 3422–3428. https://doi.org/10.1016/j.asoc.2012.12.031

Chen, F., Wang, Y., Li, M., Wu, H., & Tian, J. (2014). Principal association mining: An efficient classification approach. *Knowledge-Based Systems*, *67*, 16–25. https://doi.org/10.1016/j.knosys.2014.06.013

Chen, L., Qi, L., & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications*, *39*(10), 9588–9601. https://doi.org/10.1016/j.eswa.2012.02.158

Chen, L., & Wang, F. (2013). Preference-based clustering reviews for augmenting e-commerce recommendation. *Knowledge-Based Systems*, *50*, 44–59. https://doi.org/10.1016/j.knosys.2013.05.006

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2015). Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data. *European Journal of Operational Research*, *241*(2), 422–434. https://doi.org/10.1016/j.ejor.2014.09.008

Cheng, C. J., Chiu, S. W., Cheng, C. B., & Wu, J. Y. (2012). Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan. *Scientia Iranica*, *19*(3), 849–855. https://doi.org/10.1016/j.scient.2011.11.045

Cheung, C. F., & Li, F. L. (2012). A quantitative correlation coefficient mining method for business intelligence in small and medium enterprises of trading business. *Expert Systems with Applications*, *39*(7), 6279–6291. https://doi.org/10.1016/j.eswa.2011.10.021

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, *7*(2), 596–615.

Delgado, M. R., Mata, N. Ú. C., Yepes-Baldó, M., Montesinos, J. V. P., & Olmos, J. G. (2013). Data mining and mall users profile. *Universitas Psychologica*, *12*(1), 195–207.

Devi, B. N., Devi, Y. R., Rani, B. P., & Rao, R. R. (2012). Design and implementation of web usage mining intelligent system in the field of e-commerce. *Procedia Engineering*, *30*(2011), 20–27. https://doi.org/10.1016/j.proeng.2012.01.829

Do, N., Bae, S., & Park, C. (2015). Interactive analysis of product development experiments using On-line Analytical Mining. *Computers in Industry*, *66*, 52–62. https://doi.org/10.1016/j.compind.2014.09.003

Eduardo, L., & Vega, G. (2011). Modeling of bidding prices in power markets using clustering and fuzzy association rules, 108–117.

Espino Timón, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso, 65. Retrieved from http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memòria.pdf

Galit Shmueli, Peter C. Bruce, Mia L, N. R. P. (2014). Data Mining for Bussines Analytics-concepts, techniques, and applications whit JMP PRO. *Accv*. https://doi.org/10.1007/978-1-4614-7669-6

Gordillo-Ruiz, J. L., Martínez-Miranda, E., & Stephens, C. R. (2012). Develando estrategias de mercado: minería de datos aplicada al análisis de mercados financieros. *Computacion Y Sistemas*, *16*(2), 221–231.

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). *Principles of data mining*. *Drug safety : an international journal of medical toxicology and drug experience*

(Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

José Solano Rojas, B. (2010). Tareas de la minería de datos: clasificación CI-2352 Intr. a la minería de datos.

Khader, N., Lashier, A., & Yoon, S. W. (2016). Pharmacy robotic dispensing and planogram analysis using association rule mining with prescription data. *Expert Systems with Applications*, *57*, 296–310. https://doi.org/10.1016/j.eswa.2016.02.045

Khalifelu, Z. A., & Gharehchopogh, F. S. (2012). Comparison and evaluation of data mining techniques with algorithmic models in software cost estimation. *Procedia Technology*, *1*, 65–71. https://doi.org/10.1016/j.protcy.2012.02.013

Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications*, *57*, 311–323. https://doi.org/10.1016/j.eswa.2016.03.043

Kim, K. Y., & Lee, B. G. (2014). Marketing insights for mobile advertising and consumer segmentation in the cloud era: A Q-R hybrid methodology and practices. *Technological Forecasting and Social Change*, *91*, 78–92. https://doi.org/10.1016/j.techfore.2014.01.011

Kopaneli, A. (2014). Finance, Marketing, Management and Strategy Planning. A Qualitative Research Method Analysis of Case Studies in Business Hotels in Patras and in Athens. *Procedia Economics and Finance*, *9*(Ebeec 2013), 472–487. https://doi.org/10.1016/S2212-5671(14)00049-5

Lee, K. C., Lee, H., Lee, N., & Lim, J. (2013). An agent-based fuzzy cognitive map approach to the strategic marketing planning for industrial firms. *Industrial Marketing Management*, *42*(4), 552–563. https://doi.org/10.1016/j.indmarman.2013.03.007

Leung, C. K., MacKinnon, R. K., & Tanbeer, S. K. (2014). Tightening Upper Bounds to the Expected Support for Uncertain Frequent Pattern Mining. *Procedia Computer Science*, *35*, 328–337. https://doi.org/10.1016/j.procs.2014.08.113

Li, Y. M., Lin, C. H., & Lai, C. Y. (2010). Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications*, *9*(4), 294–304. https://doi.org/10.1016/j.elerap.2010.02.004

Liao, S. H., & Chou, S. Y. (2013). Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio. *Expert Systems with Applications*, *40*(5), 1542–1554. https://doi.org/10.1016/j.eswa.2012.08.075

Liao, S. H., Chu, P. H., Chen, Y. J., & Chang, C. C. (2012). Mining customer knowledge for exploring online group buying behavior. *Expert Systems with Applications*, *39*(3), 3708–3716. https://doi.org/10.1016/j.eswa.2011.09.066

Loshin, D. (2013). *Business Intelligence: The Savvy Manager's Guide*. *Morgan Kauf*. https://doi.org/10.1016/B978-0-12-385889-4.00001-6

Luki, J., Radenkovi, M., Despotovi-Zraki, M., Labus, A., & Bogdanovi, Z. (2016). A hybrid

approach to building a multi-dimensional business intelligence system for electricity grid operators. *Utilities Policy*, *41*, 95–106. https://doi.org/10.1016/j.jup.2016.06.010

Marín, J. (1982). Los mapas auto-organizados de Kohonen (SOM) Introducción, 1–13.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

Nafari, M., & Shahrabi, J. (2010). A temporal data mining approach for shelf-space allocation with consideration of product price. *Expert Systems with Applications*, *37*(6), 4066–4072. https://doi.org/10.1016/j.eswa.2009.11.045

Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences*, *80*(1), 57–71. https://doi.org/10.1016/j.jcss.2013.03.008

Overview, C. (n.d.). Data mining for business intelligence, *9*(Unique 03760).

Pei, J., Kamber, M., & Jiawei, H. (2012). *Data mining : concepts and techniques*.

Peng, Y., Zhang, Y., Tang, Y., & Li, S. (2011). An incident information management framework based on data integration, data mining, and multi-criteria decision making. *Decision Support Systems*, *51*(2), 316–327. https://doi.org/10.1016/j.dss.2010.11.025

Peña, A. (2006). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las organizaciones.*

Pinzon Cadena, L. L. (2011). Aplicando minería de datos al marketing educativo. *Notas D Marketing*, *1*(1), 45–61. Retrieved from http://www.usergioarboleda.edu.co/investigacion-marketing/marketing/articulo5MineriaDatos.pdf

POPEANGĂ, J., & LUNGU, I. (2012). Real-Time Business Intelligence for the Utilities Industry. *Database Systems Journal*, *3*(4), 15–24.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. *Lecture Notes for Stanford CS345A Web Mining* (Vol. 67). https://doi.org/10.1017/CBO9781139058452

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500. https://doi.org/10.1016/j.dss.2010.11.006

Resendiz Trejo, J. (2006). Las maquinas de vectores de soporte para identificación en línea.

Ríos, S. a., & Videla–Cavieres, I. F. (2014). Generating Groups of Products Using Graph Mining Techniques. *Procedia Computer Science*, *35*, 730–738. https://doi.org/10.1016/j.procs.2014.08.155

Sahoo, J., Das, A. K., & Goswami, A. (2015). An efficient approach for mining association rules from high utility itemsets. *Expert Systems with Applications*, *42*(13), 5754–5778. https://doi.org/10.1016/j.eswa.2015.02.051

Shmueli, G., Patel, N., & Bruce, P. (2007). *Data mining for business intelligence*. *Hoboken, NJ, USA*. Retrieved from http://www.c-elt.com/Data-Mining-flyer.pdf

Su, Q., & Chen, L. (2015). A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electronic Commerce Research and Applications*, *14*(1), 1–13. https://doi.org/10.1016/j.elerap.2014.10.002

Tang, H., Liao, S. S., & Sun, S. X. (2013). A prediction framework based on contextual data to support Mobile Personalized Marketing. *Decision Support Systems*, *56*(1), 234–246. https://doi.org/10.1016/j.dss.2013.06.004

Thorleuchter, D., & Van Den Poel, D. (2012). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, *39*(17), 13026–13034. https://doi.org/10.1016/j.eswa.2012.05.096

Tien, J. M. (2014). *A global view of big data*.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. *Business Intelligence: Data Mining and Optimization for Decision Making*. https://doi.org/10.1002/9780470753866

Warkentin, M., Sugumaran, V., & Sainsbury, R. (2012). The role of intelligent agents and data mining in electronic partnership management. *Expert Systems with Applications*, *39*(18), 13277–13288. https://doi.org/10.1016/j.eswa.2012.05.074

Wen, C. H., Liao, S. H., Chang, W. L., & Hsu, P. Y. (2012). Mining shopping behavior in the Taiwan luxury products market. *Expert Systems with Applications*, *39*(12), 11257–11268. https://doi.org/10.1016/j.eswa.2012.03.072

Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, *10*(3), 331–341. https://doi.org/10.1016/j.elerap.2010.11.002

Yan-li, Z., & Jia, Z. (2012). Research on Data Preprocessing In Credit Card Consuming Behavior Mining. *Energy Procedia*. https://doi.org/10.1016/j.egypro.2012.02.147

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., & Boccaletti, S. (2016). Combining complex networks and data mining: Why and how. *Physics Reports*, *635*, 1–44. https://doi.org/10.1016/j.physrep.2016.04.005

Zhang, Y., Mukherjee, R., & Soetarman, B. (2013). Concept extraction and e-commerce applications. *Electronic Commerce Research and Applications*, *12*(4), 289–296. https://doi.org/10.1016/j.elerap.2013.03.008

Zhu, Z. (2013). Discovering the influential users oriented to viral marketing based on online social networks. *Physica A: Statistical Mechanics and Its Applications*, *392*(16), 3459–3469. https://doi.org/10.1016/j.physa.2013.03.035