Universidad de Costa Rica
Sede del Pacífico
Arnoldo Ferreto Segura

Carrera

Informática y Tecnología Multimedia

Curso

Manejo de Bases de Datos

Profesor

Melber Dalorso

Investigación

Minería de Datos

Estudiantes

Jeannette Vargas Varela

B47443

Ignacio Elizondo Alvarado

B42338

Leandro Bello Delgado

B40948

Marcos Molina López

B03960

Período

II Ciclo 2017

Índice

Objetivos

Objetivo general

Revelar la función, virtudes e impactos de la minería de datos en el entorno empresarial, mediante la consulta de libros de texto y medios electrónicos para poder aumentar nuestros conocimientos sobre temas relacionados con el tema de “Minería de Datos” para el curso de Manejo de Bases de Datos.

Objetivos específicos

1. Identificar las necesidades empresariales que conlleva a la utilización de minería de datos.

2. Resumir la preparación de datos, para tener una idea de la preparación del material principal de la Minería de Datos.

3. Técnicas de minería de datos: reglas de asociación, clasificación y agrupamiento.

4. Describir las tareas y métodos de la minería de datos, para obtener una idea sobre su funcionamiento en el análisis de datos.

5. Elaborar un proyecto utilizando el lenguaje R, que ejemplifica la utilización de la minería de datos.

Introducción

El avance de la tecnología en las últimas tres décadas ha facilitado enormemente el acceso a grandes volúmenes de datos. La cantidad de información que se puede manejar hoy en día obliga a abordar el estudio de los datos/información desde una perspectiva global y no fragmentada. En los años 80 apareció el concepto de minería de datos, esta técnica se vinculó estrechamente con la dirección de empresas y en concreto al marketing.

Día a día generamos información lo que conlleva a tener una gran cantidad de la misma, lo cual implica que el generar información, puede ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones de cualquier ámbito según el dominio en el que se desarrolle.

La minería de datos no surge por la aparición de nuevas tecnologías, sino que se crea, por la aparición de nueva necesidades, los datos se transforman de productos a ser materia prima que se tiene que explotar para obtener el verdadero valor del producto, que sería en este caso el correcto uso del conocimiento que ha de ser especialmente valioso para la toma de decisiones sobre el ámbito en el que se han desarrollado recopilado o extraído los datos.

El concepto de minería de datos según witten y frank (2000). “Se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos”. Es decir la tarea fundamental de la minería de datos es encontrar relaciones a partir de los datos. Para que el proceso sea efectivo debería de ser automático o semiautomático y los patrones encontrados debería ayudar a tomar decisiones más seguras que reporten algún beneficio a la organización, es decir de una manera, es convertir datos en conocimiento, por lo tanto para ayudar al lector en conceptos de minería de

datos y sus áreas, presentamos los siguientes objetivos redactados más ampliamente.

Desarrollo

Capítulo I: Identificar las necesidades empresariales que conlleva a la implementación de minería de datos

Establecer cuál es el contexto del negocio, los objetivos del mismo y plasmarlos en objetivos de minería de datos, es previo pararse a pensar en recopilar y preparar lo datos , realizar los modelos, evaluarlos y utilizarlos. A raíz de estos es muy difícil realizar un plan de minería de datos(o cómo entender como hacer uno) sin conocer la tecnología.

La minería de datos es el proceso de recolectar y analizar grandes cantidades de datos (típicamente históricos) desde diferentes perspectivas para identificar patrones, correlaciones o tendencias ocultas entre muchas variables para posteriormente inferir las razones implícitas para estos comportamientos y proyectarlos al futuro.

Usando herramientas especializadas, los analistas de una organización pueden encontrar respuestas a complejas preguntas de negocios y de esta manera resolver retos y desafíos empresariales asociados al aumento de ingresos, la reducción en costos, u otros aspectos clave del negocio.

Sea de alguna manera u otra hoy en día ninguna empresa sea denominada grande o pequeña, puede permitir en no pensar en el tema de “ ¿implementación de minería de datos?” , cómo se trata de explicar más adelante el principal problema es no haberse hecho esa pregunta antes.

Claves del éxito de un programa de minería de datos

Se pueden destacar los siguientes aspectos fundamentales para el éxito de la minería de datos en una organización:

- El negocio y sus necesidades han de dirigir el desarrollo del programa. Se han de especificar claramente los problemas y objetivos del negocio. Con estos problemas y objetivos de negocio podremos averiguar qué datos van hacer necesarios y podrán surgir los objetivos y tareas de minería de datos.
- Una buena especificación de problemas concretos y específicos de minería de datos es otra clave del éxito, esto ayudará a establecer tareas de minería de datos que ayuden a resolver las necesidades de la organización.
- La calidad de datos es primordial.
- El uso de herramientas integradas y de entornos amigables

Uno de los aspectos a considerar cuando se utilizan programas basados en nuevas tecnologías, es que, al avanzar en el programa, la tecnología también ha progresado. De hecho hay que pensar que el programa puede ser revisado por nuevos avances que los puedan complementar.

Capítulo II: Preparación de Datos

El proceso de la preparación de datos, se dividen en tres fases y son las siguientes:

- Recopilación
- Limpieza y transformación
- Exploración y selección

A continuación se va a explicar las tres fases que se ejecutan en la “Minería de Datos”.

1. Recopilación. Almacenes de datos

Este es el primer paso para lograr obtener el objetivo de la minería de datos, que es obtener conocimiento, este conocimiento no puede ser creado si no se tiene información, ahí entra en juego la “Recopilación de datos”

Esta “Recopilación de datos” se realiza para una tarea específica, por lo tanto no se trabaja con toda la información que fluya, nos enfocamos en información necesaria para esa tarea, en sí hay que reconocer y reunir esos datos con los que se va a trabajar. Si una de las tareas no involucra gran cantidad y variedad de datos, el sentido común puede ser suficiente para obtener un conjunto de datos de calidad el cual sirva para poder empezar a trabajar. En cambio, si se necesita datos que provengan de distintas fuentes, es decir, tanto externas como internas a la organización en la cual se encuentra realizando el proceso de minería de datos, lo más seguro es que esa información sea variada y en gran cantidad, por lo tanto, el sentido común no va funcionar, no va a poder realizar tareas con esa gran cantidad de datos, porque el ser humano no está hecho para manejar grandes cantidades de información, por lo tanto no somos capaces de recopilar información necesaria a gran velocidad, es decir a como vaya llegando los datos, vamos seleccionando. De esta deficiencia que tiene el ser humano es de donde nace la idea de una tecnología

que optimice el proceso, ahí entra en juego los “Almacenes de Datos” la cual está diseñada especialmente para organizar grandes volúmenes de datos.

Necesidad de los Almacenes de Datos

En las Bases de Datos de la empresas fluye poder, es decir hay tanta información que si los empresarios que son los encargados de tomar decisiones utilizan esta información de forma productiva, la empresa va a seguir a flote y hasta puede mejorar, pero esta información se encuentra ahi guardada en sus bases, en su sistema transaccional y necesitan de un sistema analítico para poder sacarle provecho.

Para poder entender el por qué surgen los almacenes de datos, hay que conocer sobre estos dos sistemas de información (transaccional y analítico) por lo tanto a continuación se abarcara esos dos puntos.

Sistemas de información, **OLTP Y OLAP**

OLTP (On-Line Transactional Processing)

Este es el trabajo primario en un sistema de información, porque este consiste en realizar transacciones, es decir actualizaciones, consultas… en tiempo real dentro de la base de datos, con un objetivo operacional: hacer funcionar las aplicaciones de la organización, proporcionar información sobre el estado del sistema de información y permitir actualizarlo, un ejemplo de este trabajo transaccional son, insertar un nuevo cliente, hacer una modificación en el salario de algún empleado, el trámite de un pedido, almacenar una venta… Estos procesos son trabajos diarios de la base de datos, para lo que fue creada inicialmente. En sí, eso es un sistema transaccional.

OLAP (On-Line Analytical Processing)

Este otro sistema de información, también es en tiempo real. Este engloba varias operaciones, que ayuden a cruzar gran cantidad de información; el objetivo

de esto es realizar consultas para obtener informes y resúmenes, los cuales se utilizan como apoyo en las tomas de decisiones, por ejemplo, un resúmen de las ventas mensuales, el producto que más se ha vendido en el último trimestre… Son muchos los análisis que se pueden realizar, y los departamentos que más lo requieren son los de dirección, logística, por ejemplo.

Una característica de ambos procesamientos es que sean “on-line” que sean “instantáneos” es decir que se puedan realizar en cualquier momento (en tiempo real)

Una pregunta que podría surgir es ¿por qué ambos procesamientos no se realizan en una misma base de datos transaccional si estas se encuentran normalizadas por lo que va a evitar información redundante lo cual evita análisis en vano?

Sucede que si los dos procesamientos se llevan a cabo en una misma base de datos y que ésta sea transaccional, va a ocasionar dos problemas:

- Las consultas OLAP desordenan o perturban el trabajo transaccional diario, porque este procesamiento realiza consultas complejas, que van a involucrar muchas tablas y esto consume gran parte de los recursos del sistema de gestión de las bases de datos y como resultado es que durante la ejecución de esas consultas va a afectar las operaciones transaccionales normales del OLTP, va a provocar que las aplicaciones vayan más lentas, las actualizaciones se demoren y el sistema podría llegar a colapsar.
- La base de datos está hecha para el trabajo transaccional y no para analizar datos, por lo tanto si se tuviera el sistema dedicado para realizar una consulta OLAP, esta consulta podría llegar a necesitar mucho tiempo, pero no solo por ser compleja sino porque el esquema de la base de datos no es la adecuada para ese tipo de consultas asesinas. Se le llaman consultas asesinas (killer queries) porque este proceso puede llegar a provocar el colapso en el sistema, el que fue mencionado en el primer problema.

En sí, es imposible un análisis complejo de la información en tiempo real si ambos procedimientos se realizan sobre la misma base de datos, si se realizan consultas OLAP de esa forma, deberían esperar las noches o fines de semanas, por lo tanto dejarían de ser “on-line” porque no se están ejecutando en tiempo real, y la información que se analiza ya podría estar desactualizada. Es por estas razones que se separa las bases de datos transaccionales con las bases de datos analíticas y esto genera o hace que nazcan los “Almacenes de datos” como una herramienta para mejorar el proceso de análisis, porque estaría diseñada solo para esa función, aunque existen otras formas, en este trabajo se enfoca en el “Almacén de Datos” porque es eficiente.

Un Almacén de Datos, es una “colección de datos que ayuda a tomar decisiones” pero no son imprescindibles para hacer extracción de conocimiento a partir de datos, porque una minería de datos se puede hacer sobre un archivo de datos, sin embargo, es una ventaja tener esta organización de un almacén, tanto a medio como largo plazo, porque los datos van en aumento; además, un almacén también tiene sentido si la información no proviene de una base de datos transaccional. En gran medida un Almacén de Datos también facilita la limpieza y la transformación de datos, pero ese proceso de limpieza y transformación forma parte de otro de los procesos de la “Preparación de Datos” que a continuación se abarca. (Hernández, J., Ramírez, J & Ferri, C. (2005)).

2. Limpieza y transformación

La limpieza es otro proceso de preparación de datos que surge después de la recopilación de los datos, y esto se genera para que esta información esté en condiciones para ser analizadas. Los beneficios que genere el análisis de información, es decir los conocimiento que se generan dependen gran parte de la calidad de los datos recopilados, por eso es que realizan estos procedimientos o filtros.

Una vez con datos ya recopilados en el proceso de limpieza se van a encontrar problemas en la calidad de esos datos y esto se puede empeorar con información que llega de otras fuentes, es decir no es del proceso transaccional de la empresa, uno de los problemas que puede surgir es que no llegue toda la información, por lo que en algunos casos sí hay y en otros no, por ejemplo, si se recolecta información sobre los empleados de otra empresa porque también me puede servir esa información, pero el atributo “Género” en algunos casos se encuentra vacío, no se puede tener información exacta de cuántos hombres o cuántas mujeres hay en esa empresa, por decir algo. Estos datos no tiene una solución sencilla pero la información duplicada sí, porque tiene que ser identificada desde el momento de integración de datos, porque la Base de Datos tiene que estar normalizada.

El proceso de integración se realiza durante la recopilación de datos y durante el proceso se mantiene un almacén de datos, en ese momento es cuando la limpieza de datos puede en muchos casos detectar y solucionar problemas en los datos que se están integrando, como en los datos faltantes, entonces es un proceso lógico, porque la limpieza se realiza en el momento de la integración o inmediatamente después de ella.

Para entender un poco el término “Integración” se va a explicar a continuación…

Integración:

Aquí ocurre un problema a la hora de la integración de datos que provienen de distintas fuentes, el problema es conseguir que los datos se unifiquen para los datos que se parecen y conseguir que se separen para los datos diferentes. Este problema se conoce como “El problema de esclarecimiento de la identidad” por ejemplo:

IDENTIFICADORES: se realizan con identificadores externos a la Base de Datos, como, número de identidad, número de póliza, matrículas, tarjeta de crédito…

nota: se suele ser conservador a la hora de identificar, si no se está seguro no se hace, esta tarea puede ser difícil, ya que si se utilizan claves internas para identificar, hay que mirar los identificadores externos y estos muchas veces varían de formato, por ejemplo:

Un ciudadano español se puede identificar con el DNI, el NIF y el pasaporte, y estos coinciden en 8 dígitos, pero el NIF añade una letra y el pasaporte añade alguna letra y dígito adicional, un ejemplo de esto sería:

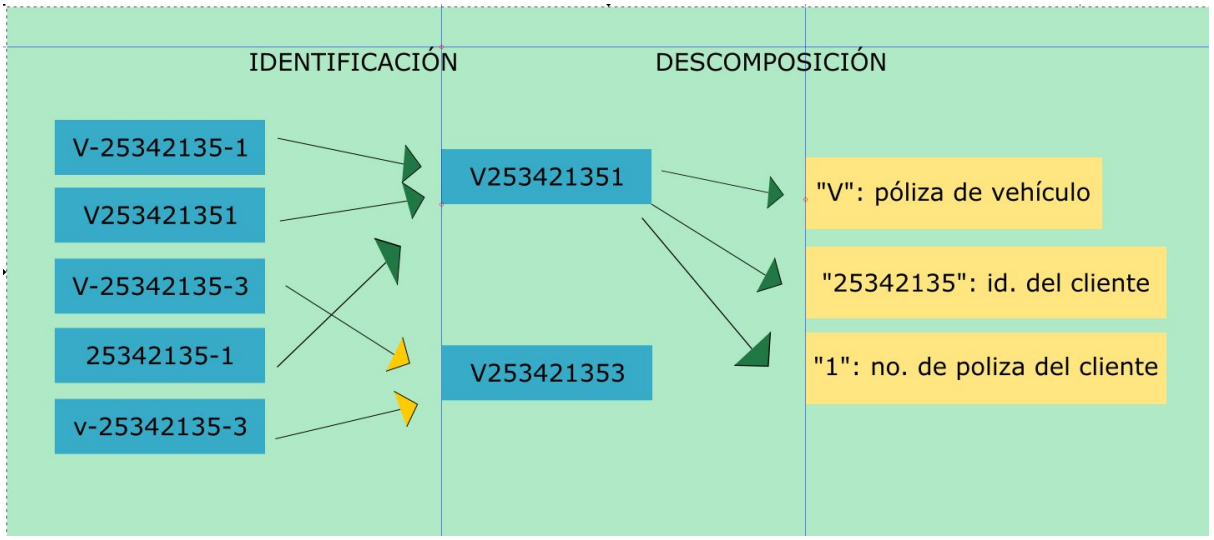

Figura 1: Ejemplo de integración: identificación y descomposición

Otro problema durante la integración son los datos duplicados, esto provocado por claves internas mal diseñadas, las cuales tiene que ser identificadas en el momento de la integración y se le denomina “Descomposición de Claves”.

Además, cuando se integran de forma correcta dos fuentes diferentes de datos de distintos objetos suele suceder que puedan aparecer datos faltantes (se registra en una fuente pero no en otra) o datos inconsistentes (datos diferentes en una fuente y otra)

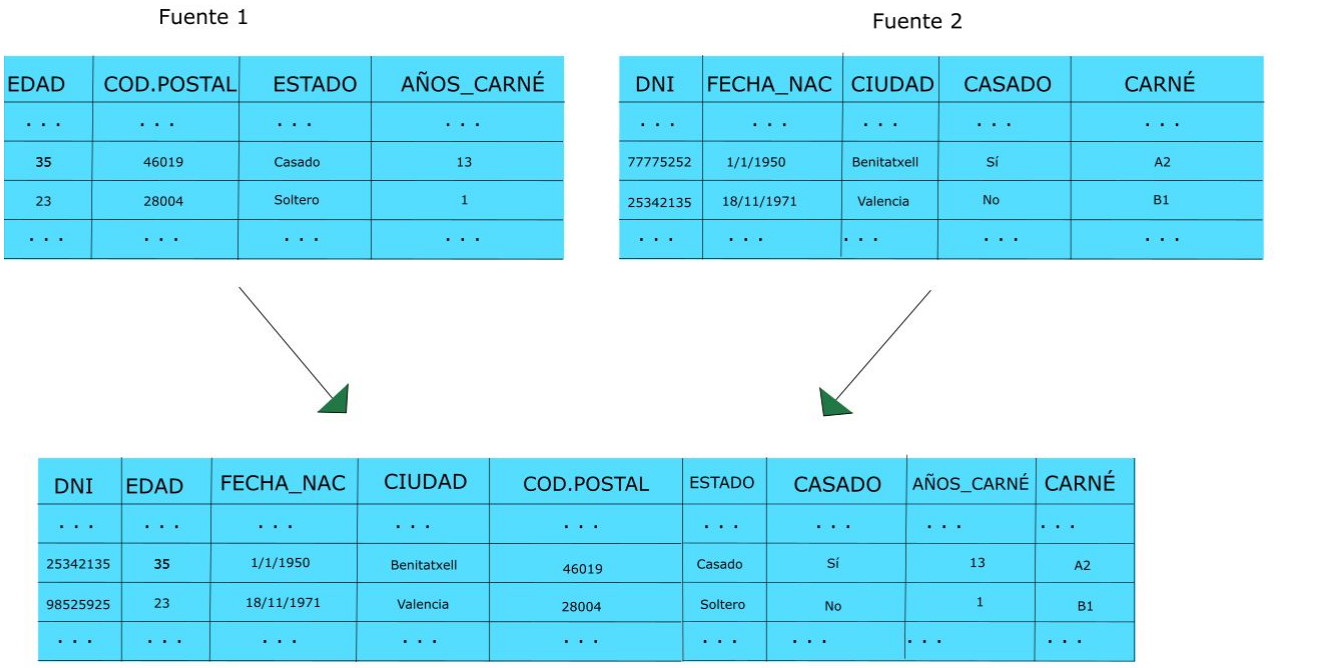

Figura 2: Ejemplo de integración de atributos de distintas fuentes

Aparecen campos redundantes como: “edad” y “fecha_nac”, “ciudad” y “cod_postal”, “estado” y “casado”. Por lo tanto se tratan de fusionar, aplicando normalización. En muchos casos los datos inconsistentes (datos diferentes de una fuente y de otra) se convierten en faltantes, por ejemplo si el mismo cliente tiene “estados civiles” diferentes en cada fuente, es mejor dejar el valor sin nada, para no tener que elegi al azar.

Cuando ya se tiene integrados los datos se realiza un “Reconocimiento” el cual se explica a continuación:

Reconocimiento:

Cuando se tiene los datos integrados se comienza con el “Reconocimiento” y consiste en realizar un resúmen de las características, ya sea tabla por tabla o para toda la base o almacén de datos, por ejemplo:

Para una Compañía de Seguros, tenemos datos que hacen referencia a las “Pólizas de Vehículos”. La siguiente tabla muestra un resumen incompleto de los atributos de esa base de datos:

Atributo	Tabla	Tipo	# total	# nulos	# dists	Media	Desv.e.	Moda	Min	Max
Código postal	Cliente	Nominal	10320	150	1672	-	-	“46003”	“01001”	“50312”
Sexo	Cliente	Nominal	10320	23	6	-	-	“V”	“E”	“M”
Estado civil	Cliente	Nominal	10320	317	8	-	-	Casado	“Casado”	“Viudo”
Edad	Cliente	Numérico	10320	4	66	42,3	12,5	37	18	87
Total póliza p/a	Póliza	Numérico	17523	1325	142	737,24€	327€	680€	375€	6200€
Asegurados	Póliza	Numérico	17523	0	7	1,31	0,25	1	0	10
Matrícula	Vehículo	Nominal	16324	0	16324	-	-	-	“A-0003-BF”	“Z-9835-AF”
Modelo	Vehículo	Nominal	16324	1321	2429	-	-	“O. Astra”	“Audi A3”	“VW Polo”
...	...	...	...	...	...	...	...	...	...	...

Figura 3: Tabla o resumen de atributos

La tabla anterior se puede construir con consultas SQL y da gran cantidad de información de un simple vistazo. Además de ver cuántos clientes, pólizas y vehículos se tiene, se puede observar el total de “nulos” de cada atributo, la “moda” que es el valor más frecuente.

Un dato que resalta a la vista que tiene que ver con la calidad de los datos, por ejemplo:

¿Cómo va a ver varios datos para identificar el sexo de una persona?

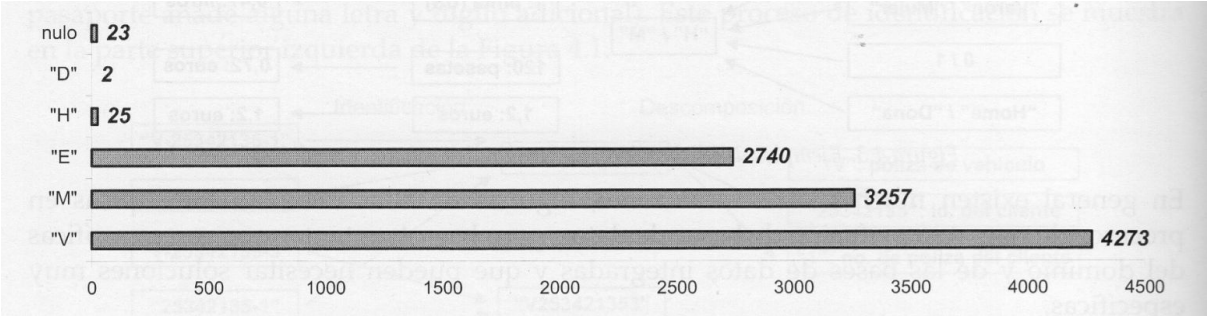

Figura 4: Histograma representando las frecuencias de un atributo

Al observar los datos nos podemos dar cuenta que el nombre “sexo” es un atributo mal elegido, porque el valor “E” representa a una empresa que asegura su vehículo, es decir se sabe que no es una persona. Los datos “V” , “M” y “E” parecen ser claros, “varón”, “mujer” y “empresa” el problema se encuentra en los valores “H” y “D” . Entonces realizan un análisis de donde proviene los datos y se descubre que la mayoría de “H” se supone que representa a los “Hombres” pero que algunos pueden venir de forma incorrecta del término erróneo “Hembra” especialmente en datos antiguos, que se encontraban en papel y no de una aplicación informática que tiene mayores restricciones de integridad. Por lo tanto, debido a su antigüedad, no es una información confiable, entonces se deja como valor nulo (más adelante se puede rellenar muchos de esos valores mirando los nombres de los clientes). El valor “D” podría ser que las aplicaciones de la empresa son bilingües (castellano/ catalán) y la “D” es para representar a la “Mujer” por en catalán “Dona” significa “Mujer”. Este es un proceso delicado que puede cambiar un análisis.

Luego de “Integrar” y “Reconocer” los datos se realiza una transformación de ellos, para reducir la dimensionalidad porque suele ser alta, y esto puede ser un problema a la hora de aprender de los datos, si se tiene mucha dimensionalidad es decir, muchos atributos puede ser un toque complicado encontrar patrones, hay técnicas que ayuda en la transformación pero no se contemplan en esta investigación. (Hernández, J., Ramírez, J & Ferri, C. (2005))

3. Exploración y Selección

Una vez que se tiene recopilados y limpios los datos, todavía no se encuentra la información lista para aplicar una tarea de minería de datos, se tiene que realizar un reconocimiento en ellos, es decir realizar un análisis exploratorio para conocerlos mejor relacionándolos con la tarea de minería y seleccionarlos.

Se explicará este proceso por medio de un ejemplo:

Imagínese que le cae del cielo una Base de Datos o Almacén de Datos con una nota que dice, “extraiga usted conocimiento de aquí” Aparte de la sorpresa natural de ver llover Bases de Datos, que achacará posiblemente al cambio climático, usted debería preguntarse lo siguiente:

- ¿Qué parte de los datos es pertinente analizar?
- ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
- ¿Qué conocimiento puede ser válido, novedoso e interesante?
- ¿Qué conocimiento previo me hace falta para realizar esta tarea?

Éstas pregunta son para tener claro el conocimiento que se necesita, si usted no tiene claro estos puntos, por ejemplo, no va a poder aprovechar los datos y el conocimiento es nulo o erróneo, porque una herramienta de Minería de Datos no puede digerir datos y crear un conocimiento razonable, porque no es capaz de pensar por sí mismo, necesita de la inteligencia humana para generar los análisis, en el sentido de construir el camino que se quiere minar. En sí, hay que conocer qué conocimiento es útil, eso es lo primero que hay que tener claro, porque si no, no se puede decidir que parte de los datos lo pueden proporcionar y se vuelve ineficiente la minería. (Hernández, J., Ramírez, J & Ferri, C. (2005))

Capítulo III: Técnicas de minería de datos: reglas de asociación, clasificación y agrupamiento.

Es común relacionar el término *conocimiento* como algo que implica cierto grado de inteligencia. En este sentido, el avance desde los simples datos a la información y después al descubrimiento de conocimiento a medida que se aplica más procesamiento a los datos. Se clasifica el conocimiento como algo inductivo en lugar de algo deductivo.

Con respecto a la información, Elmasri & Shamkant (2002), manifiestan:

> La minería de datos trata con el **conocimiento inductivo,** mediante el cual se descubren nuevas reglas y patrones a partir de los datos suministrados. El conocimiento se puede representar de varias formas. En un sentido no estructurado, se puede representar mediante reglas o lógica proposicional. En forma estructurada, se puede representar mediante árboles de decisión, redes semánticas, redes neuronales o jerarquías de clases o marcos. (p. 826).

Es muy común, describir el conocimiento adquirido en el proceso de minería de datos según las siguientes clasificaciones:

Reglas de asociación

El modelo de la cesta de la compra, soporte y confianza

En este caso, la cesta de la compra se corresponde con el conjunto de productos que compra un consumidor en un supermercado durante su visita. Una de las principales tecnologías dentro de la minería de datos tiene relación con el

descubrimiento de reglas de asociación. La base de datos se ve como un conjunto de transacciones, cada una de ellas relacionada con un conjunto de elementos.

Esta asociación significa que si un cliente compra *X*, también comprará probablemente *Y*. Para que una regla resulte interesante deberá cumplir con alguna medida de interés. Dos medidas de interés habituales son el soporte y la confianza.

El **soporte** se calcula con respecto al conjunto; indica con qué frecuencia aparece determinado conjunto. O en otras palabras, qué porcentaje de veces es agregado a la cesta de compra una lista igual de productos . Si el soporte es bajo, se dice que no hay evidencia de que el conjunto de elementos aparezcan juntos, ya que aparece en un pequeño número de transacciones.

La **confianza** podemos considerarla como la probabilidad de que los elementos (productos del súper mercado) sean comprados en el supuesto de que la cesta de compras sea adquirida por un cliente.

Algunos algoritmos utilizados para el propósito mencionado se listan a continuación:

- El Algoritmo Apriori.
- Algoritmo de muestreo.
- Algoritmo de árbol de patrón frecuente.
- Algoritmo de particionado.

Otros tipos de reglas de asociación

Las **reglas de asociación entre jerarquías** son tipos de asociaciones que son particularmente interesantes por un motivo especial. Estas asociaciones se producen entre jerarquías de elementos. Por lo general, es posible dividir los elementos en jerarquías disjuntas, basándose en la naturaleza del dominio.

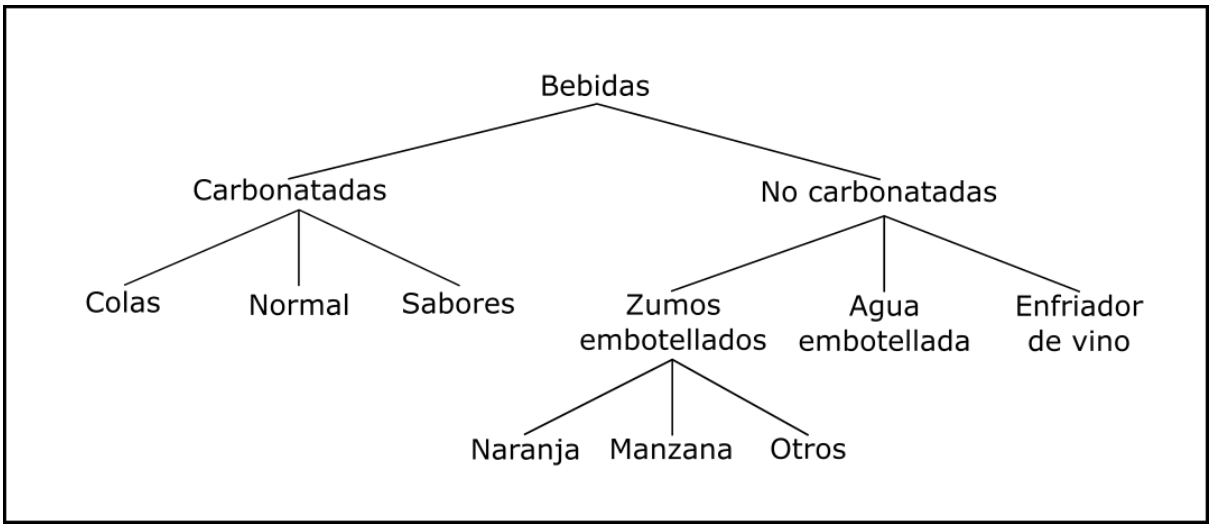

Figura 1: Asociación entre jerarquías.

Las **asociaciones multidimensionales** representan atributos de registros de un archivo o, en términos de relaciones, las columnas de las filas de una. Implica la búsqueda de patrones en un archivo.

Clasificación

Además, otra técnica existente, con igual importancia que la anterior tiene como nombre **clasificación.** Este procedimiento de minería de datos tiene como objetivo diferenciar ventajas y desventajas realizando un proceso de clasificación para determinados elementos en una colección de datos existente.

Por consiguiente, “la clasificación es el proceso de aprendizaje de un modelo que describe diferentes clases de datos. Las clases están predeterminadas” (Elmasri & Shamkant, 2002, p. 836).

Por ejemplo, en un sistema de salud es posible clasificar las personas propensas a sufrir enfermedades cardiacas de acuerdo a determinados factores.

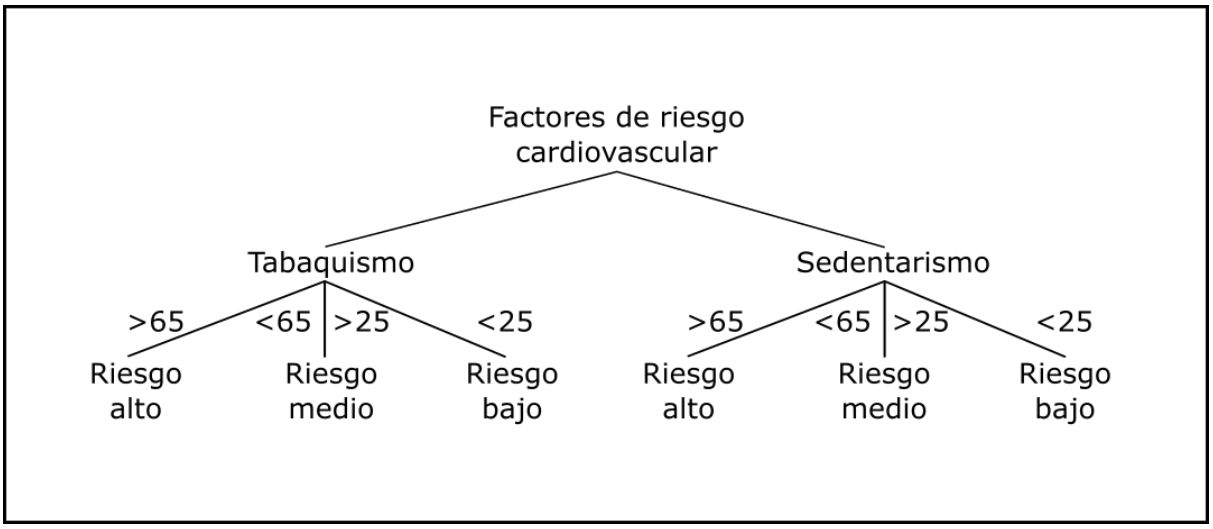

Figura 2: Clasificación..

El primer paso a seguir comienza utilizando un conjunto de datos de entrenamiento previamente seleccionados. Cada registro tiene un atributo, en concreto llamado etiqueta de clase, que indica a qué clase pertenece el registro.

El modelo resultante, tiene por lo general la forma de un árbol. Un punto importante tiene que ver con la capacidad del modelo para predecir la clasificación de los nuevos datos.

Dentro del campo de la minería de datos a este tipo de técnica de minería también se le ha llegado a conocer como aprendizaje supervisado.

Agrupamiento

El tipo anterior de minería conocido como clasificación realiza el proceso del descubrimiento de información basándose en una muestra de ejemplo clasificado previamente. No obstante, en algunos caso es bueno llegar al descubrimiento de información sin disponer de una muestra de datos clasificado previamente.

Cuando de agrupamiento se trata, al proceso descrito se le conoce también como aprendizaje no supervisado.

“El objetivo del agrupamiento es situar los registros en grupos, de forma que los registros de un grupo sean similares a los demás y distintos a los registros de otros grupos” (Elmasri & Shamkant, 2002, p. 840).

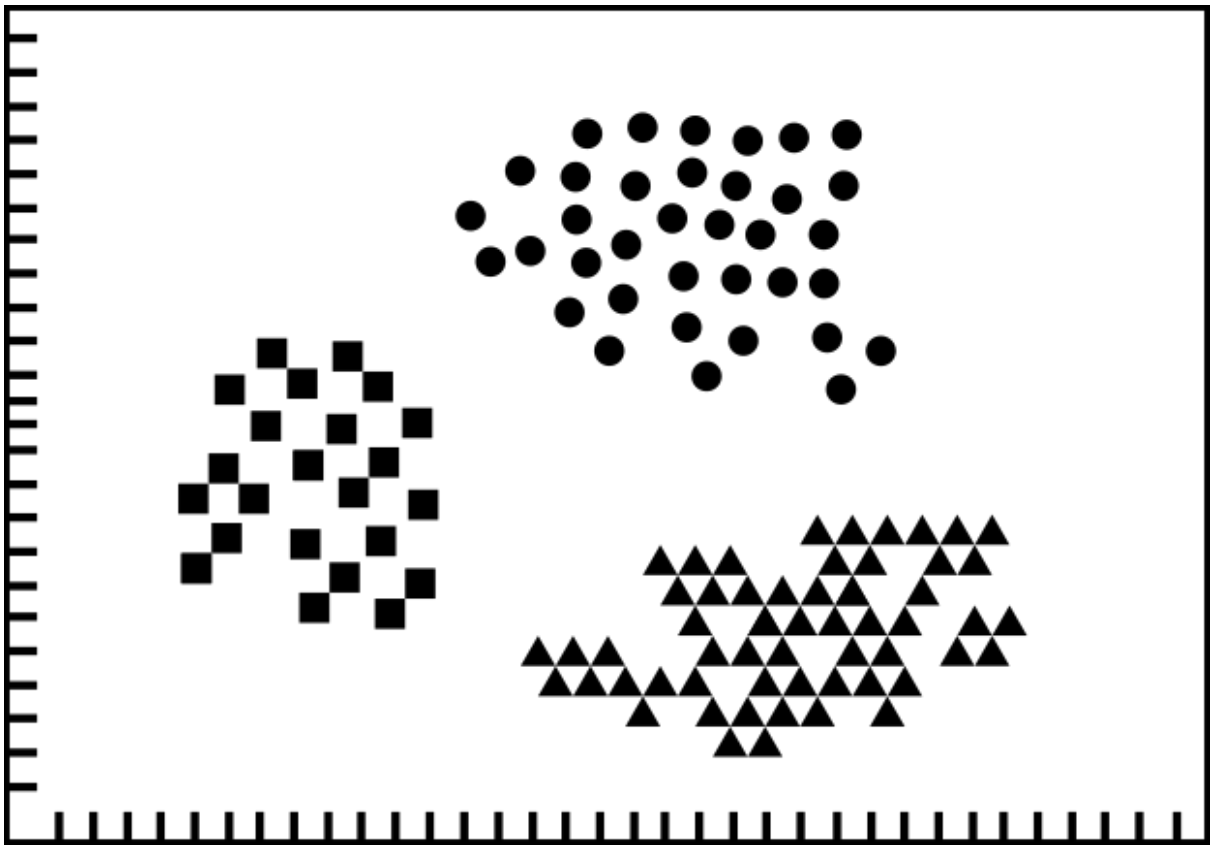

Figura 3: Agrupamiento

Otras peculiaridades acerca de la minería de datos

Redes neuronales

Es debido al gran avance de las investigaciones en inteligencia artificial que hoy en día las aplicaciones se han extendido a diversos campos de las tecnologías de información. El caso más relacionado es el de las redes neuronales.

“Una red neuronal es una técnica derivada de la investigación en inteligencia artificial que utiliza un método iterativo para llevarla a cabo. Las redes neuronales usan un modelo de ajuste de curvas para deducir una función a partir de un conjunto de muestras.” (Elmasri & Shamkant, 2002, p. 843).

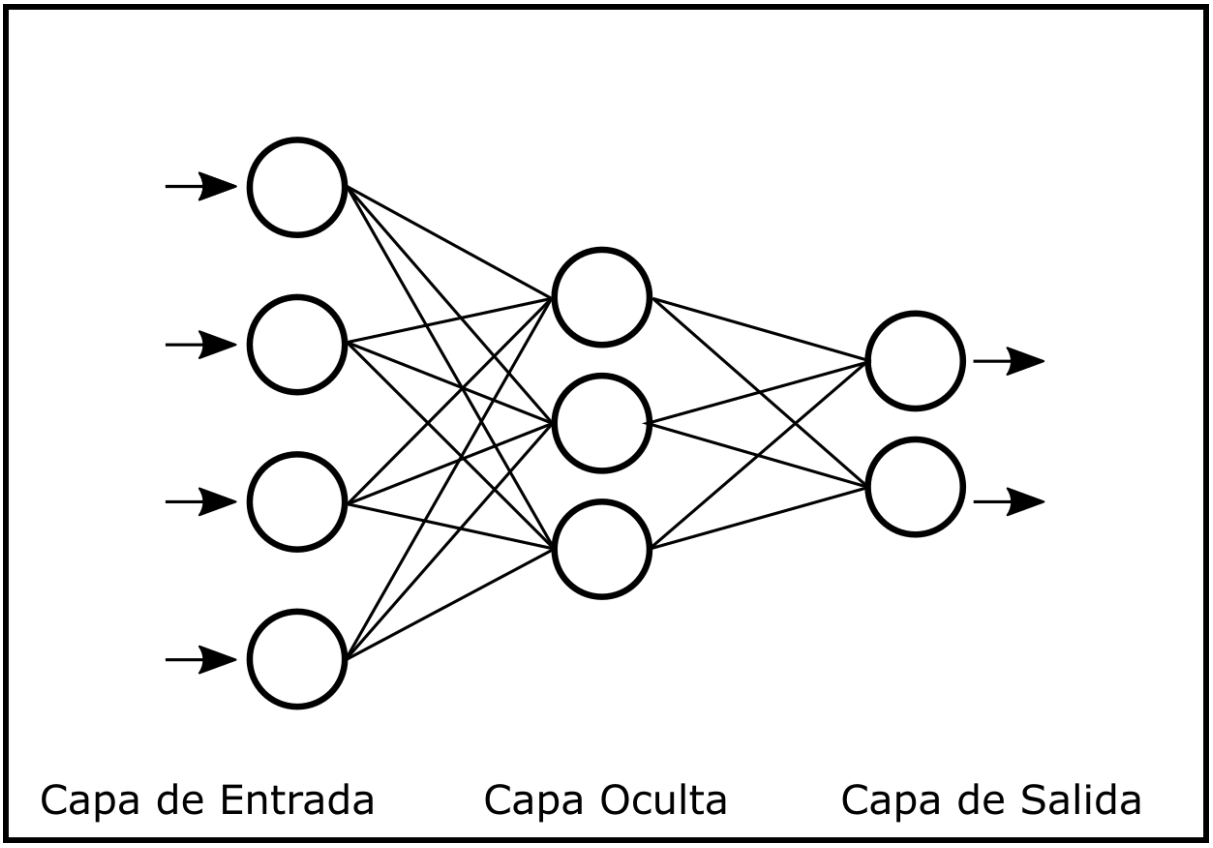

Figura 3: Redes neuronales

Las redes neuronales realizan autoadaptaciones; es decir, aprenden a partir de la información de la información existente sobre un problema determinado. Se ejecutan con efectividad en tareas de clasificación y se usan, por tanto, en la minería de datos.

Capítulo IV: Funcionamiento de tareas y Métodos de Datos

En primera instancia el objetivo que destaca en el ámbito de la minería de datos es el descubrir patrones que sean de interés, es decir válidos, comprensibles, novedosos e interesantes, esto debido a la capacidad de especies como el ser humano al observar y analizar patrones de información que cohabita en su entorno e incluso la ausencia de estos mismos.

En este capítulo se enfoca en ver las técnicas, métodos o bien el funcionamiento, esto mientras analizamos que tienen todas ellas en común y cuáles son los problemas a las que son sometidas estas técnicas.

Antes de iniciar un proceso de minería de datos debemos tener en cuenta los procesos de preparación de datos como se refiere en el capítulo número tres, el cual guíe al la transformación de un conjunto de datos en un conocimiento valioso.

A la hora de procesar una tarea esta dependerá de diversos aspectos como **bias** las cuales influyen a la al momento de expresar o bien definir patrones, donde HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005) exponen la gran diferencia en una regresión lineal a una regresión realizada por una red neuronal multicapa. Ambas pueden realizar una tarea en concreto, sin embargo la segunda **bia** anteriormente mencionada puede ayudar a refinar mucho más la búsqueda con un poco de espera para obtener el modelo.

Tareas y métodos:

El la minería de datos encontraremos términos de Tareas y Métodos los cuales debemos tener el cuidado de poder discernir, donde una tarea en este ámbito es un tipo de problema en la minería de datos; dando como ejemplo la clasificación de piezas de un proveedor X, en diferentes estados como puede ser:

óptimas, defectuosas, reparables o bien defectuosas irreparables. Dicha tarea se describe como clasificatoria pudiendo resolver mediante árboles de decisión o redes neuronales.

Tareas

En el capítulo tres se pudo apreciar algunas de las tareas más relevantes como; clasificación, regresión, agrupamiento, reglas de asociación sin embargo en este apartado se mostrará relaciones de mas complejas con la finalidad de tener una compresión un poco más acertada sobre el fundamento en las tareas en minería de datos, HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005) citan:

> Definamos *E* como el conjunto de todos los posibles elementos de entrada. Las instancias posibles dentro de *E* generalmente se representan como un conjunto de valores para una serie de atributos(sean nominales o numéricos). Es decir $E = A1 \times A2 \times ....An.$ *y un ejemplo e es una tupla* $<a1, a2, ..., an>$ *tal que* $a1 \in A1$.(p.139).

Lo cierto es que la minería de dato posee un gran número técnicas, términos, clases y fórmulas científicas donde se dividen en dos grandes vertientes llamados métodos *Predictivos* o métodos *Descriptivos*, es por este motivo que se explicará a grandes rasgos de qué trata cada uno de ellos, sin caer en la profundización debido a la limitación del trabajo de investigación.

Predictiva

Es un área de la minería de datos conocidos como métodos de aprendizaje supervisado, asimétricos o bien directos. Estos se basa en entrenar a un modelo o bien un método el cual busca predecir una o más variables, tendencias o patrones de comportamiento en relación a las demás partiendo de estos mismo datos, de este modo responde a preguntas futuras en base de un análisis de su comportamiento. Cabe destacar que estos métodos son desarrollados el ámbito de

máquina de aprendizaje como redes neuronales mediante perceptrón de multicapas y árboles de decisión.

Para conceptualizar mejor este método podemos responder preguntas como: ¿Qué tantos se podrá vender el próximo año con un producto X? ¿Qué clase de clientes comprará un mayor volúmen de un producto Y? ¿Que tipo de clientes están en riesgo de disminuir en una empresa?

Los tareas predictivas pueden clasificarse en los siguientes:

Clasificación o discriminación: Son representados como un conjunto de pares de elementos de dos conjuntos, el cual se puede expresar de la siguiente manera, $\boldsymbol{\delta}=\{ <e , s> : e \in E, s \in S \}$ donde ***S*** es descrito como el conjunto de valores de salida, respecto a la variable ***e***, al ir acompañado de un valor ***S*** es denominado ejemplos etiquetados $<e , s>$ y como consecuencia $\boldsymbol{\delta}$ es denominado en conjunto de datos.

Esta función $\lambda: E \rightarrow S$ es denominada clasificador donde para cada valor de *E* poseemos un único valor para *S* donde este también es nominal ($b_1, b_{1,} \ldots b_m$) cómo correspondencia existente, esto *según:* HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005).

La capacidad de dicha función en síntesis determinará la clase para cada nuevo ejemplo sin etiquetar; en otras palabras dará un valor de ***S*** para cada valor ***e.*** Una forma de poner en práctica dicha afirmación es con ejemplos de la siguiente naturaleza: Determinar cuál medicamento es mejor para una patología, clasificar a los estudiantes en A, B, C, D o F según sus notas, utilizando simplemente límites (60, 70, 80, 90), o bien la clasificación de correos electrónicos como spam o no. .

Clasificación suave: De la misma manera que el primer método se basa en pares de elementos de dos conjuntos $\lambda: E \rightarrow S$ para solucionar problemas de la misma clase*, sin embargo esta clasificación también cuenta con una nueva función*

θ: E ➡ ℜ el cual hace referencia al grado de certeza de una predicción realizada por la función λ.

Lo interesante de este método es que permite realizar rankings de predicciones. Si tuviéramos que mencionar uno de los ejemplos anteriores como el caso de clasificar cuál fármaco es más efectivo para determinada patología, proporcionando además la certeza de la clasificación.

Estimación de probabilidad de clasificación: se considera una generalización de la clasificación suave, sin embargo en esta función se enfoca en aprender ***m*** funciones $\theta_i: E \rightarrow \Re$ donde ***m*** es el número de clases, en otras palabras para cada función aprender ***m*** retorna un valor real p_i donde p_i es la probabilidad de la clase **i**.

Categorización: A diferencia de la Clasificación que busca de un ejemplo $\delta=\{ <e , s> : e \in E, s \in S \}$ aprender una función determinada de (una a una) expresamente en $\lambda: E \rightarrow S$ la Categorización más bien busca una correspondencia la cual nos dice que podemos designar diversas categorías a un mismo e, y al final mostrar las mejores categorías que supere un cierto valor.

Debe hacerse notar que el método de categorización puede presentarse también en categorización suave donde la categoría asignada va acompañada de su certeza, o bien una estimador de probabilidades este según explican los autores HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005).

Preferencias o priorización: Las tareas de este tipo determina dos o más ejemplos, en un orden o preferencia de ahí su nombre, sin embargo cada ejemplo en realidad es una secuencia descrita de la siguiente forma $<e_1 , e_2, ...e_k,>: e_i, \in E \geq 2$ cuyo orden de secuencia representa la predicción, de esta manera al analizar

un conjunto de de datos para este problema en específico como se expresa $\delta:\{ <e_1, e_2 ,... e_k> : e_i \in E\}$ en realidad estamos observando un conjunto de secuencias.

Regresión: Considerada la tarea más simple a definir en las tareas de predicción y es debido al conjunto de evidencias que la compone el cual que es una correspondencia entre dos conjuntos descrita como $\delta:E \rightarrow S$, donde los valores de salida es S, y por otro lado δ es nominado el conjunto de datos etiquetados. En este caso el valor para cada valor de E poseemos un valor para S.

Descriptivas:

A diferencia de las tareas Predictivas que busca predecir datos, tendencias entre otros, las tareas de carácter Descriptivas se enfocan en describir los datos existentes y convertirlos en información relevante mediante graficación, existen muchos caminos para resolver una tarea de esta naturaleza, algunas de las tareas descritas en el Capítulo Tres son consideradas tareas de tipo descriptivas, no obstante en este apartado describiremos tareas más específicas como a continuación.

Agrupamiento(clustering): Clustering juega un papel muy importante en aplicaciones de minería de datos tales como el análisis o exploración de datos de carácter científico, el proceso consiste en la división de datos en grupos de objetos similares mediante la información que brinda sus variables, cabe destacar que es muy similar a la clasificación de la vertiente predictiva, sin embargo Clustering se diferencia en que los valores de *S (*en un conjunto de elemento) y sus miembros, se crean durante el proceso de aprendizaje.

Lo práctico de los resultados de agrupamiento radica en que la información a mostrar se simplificada en gran medida a cambio de perder algunos detalles en los datos. La aplicación de este cubre bases de datos especializadas cubriendo el campo de la astronomía, también lo vemos marketing, Web, diagnósticos médicos, biología computacional, análisis del ADN y la lista sigue en muchos campos.

Correlaciones y factorizaciones: En primera instancia se enfocan en atributo de carácter numérico donde el objetivo es mostrar redundancias o dependencia entre los atributos, esta tarea se basa en ver si dos ejemplos de un conjunto $E=A_1 \times A_2 \times \ldots \times A_n$ están correlacionados linealmente o bien relacionados de algún otro modo.

Reglas de Asociación: Conocido también como análisis de asociación, es una de las tareas de mayor uso en la minería de datos, este posee el mismo objetivo de las tareas de correlación y factorización, pero para los atributos de carácter nominal. Dados dos ejemplos del conjunto E una regla de asociación se define generalmente de la forma “si A1=v1 y A2=v2 y … y Ak=vk entonces Ar=vr y As=vs y … y Az=vz” donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributo.

Dependencias Funcionales: Este tipo de tareas toman en consideración todos los posibles valores, puede ser descrita de la siguiente forma:” dados los valores de Ai, Aj, … , Ak” de esta manera se puede determinar un valor como Ar. Un ejemplo que podría ilustrar la mecánica de dicha tarea es; “dada la edad (discretizada en seis intervalos), el código postal si está casado o no“ con estos datos es posible determinar con alta seguridad si un cliente posee un vehículo, esto debido a que los atributos bando puede ser orientado o no orientado como si se tratase de las reglas de asociación esto según exponen HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005).

Detección de Valores e instancias anómalas: Esta tarea posee el objetivo de encontrar instancias que no son similares de este modo tratando de detectar comportamientos de tipo anómalos o diferenciado, el cual es implementado para detectar posibles fraudes, intrusos o bien fallos.

Hasta este momento hemos presenciado problemas de gran variedad tanto en tareas Predictivas como en Descriptivas, sin embargo existen casos especiales de problemas donde en su clasificación posee más de dos clases (multiclasificación) donde necesitamos resolverlo mediante un clasificador que solo consiga discriminar entre dos clases (clasificación binaria) es por ello que en la minería implementa un

proceso para adaptar problemas de múltiples clases en problemas de dos clases el cual es denominado **"binarización"** los cuales se clasifican en:

- Uno frente a uno.
- Todos los pares.
- Todas las mitades.

Uno frente a uno: El primer punto de binarización nos dice que es necesario la construcción de un clasificador binario, este emplea todos los ejemplos de una clase y agrupando en una misma clase el resto de ejemplos.

Todos los pares: Es utilizado por pares de clases, en este caso se da la construcción de un clasificador binario, el cual utiliza cada uno de los ejemplos de Dos clases e ignora el resto, este es descrito como n(n-1)/2 donde posteriormente acontece a combinar

Todas las mitades: Hay que destacar que la combinación de este caso es más compleja, en primera instancia se construye un clasificador binario utilizando los ejemplos de la mitad de las clases por un lado y el resto por el otro, esto quiere decir que si n clases y este número es par, tendremos n/2 por un lado y n/2 clases por el otro. Esto se lleva a cabo para toda las posibles particiones, posean o no el mismo número de clases, dando como resultado un gran incremento en el números de variables.

Métodos:

Como en cualquier problema es necesario la intervención de un método, técnicas o algoritmos para su solución, como menciona HERNÁNDEZ ORALLO, J, RAMÍREZ QUINTANA M.J. y FERRI RAMÍREZ, C. (2005). una de las cosas que

impacta más a los aprendices de la minería de datos aparte de que una tarea puede tener muchos métodos diferentes para resolverla, es que de la misma forma, un método puede resolver gran ámbito de tareas. De esta manera daremos inicio a una breve explicación de las técnicas o bien métodos para ejecutar las tareas mencionadas anteriormente.

1. **Técnicas algebraicas y estadísticas:**
2. **Técicas bayesianas:**
3. **Técnicas basadas en conteos de frecuencias y tablas de contingencia**
4. **Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas:**
5. **Técnicas relacionales,declarativas y estructurales**
6. **Técnicas basadas en redes neuronales artificiales.Técnicas basadas en núcleo y máquina de soporte vectorial:**
7. **Técnicas estocásticas y difusa:**
8. **Técnicas basadas en casos, en densidad o distancia:**

Conclusión

Las técnicas para la extracción de conocimiento en el ámbito de la minería de datos destacan por su peculiar sistema de asociación, clasificación y agrupamiento. Es debido a su potencia que hoy podemos obtener valiosa información de los datos.

Exactamente, se hace referencia a conseguir datos de los datos. Es ésta la manera de poder sacar conclusiones a partir de la implementación de técnicas tan importante como las mencionadas en el capítulo III del presente documento.

Sin duda la minería de datos es una excelente herramienta para el análisis de datos, es por ello que en los próximos años; grandes, medianas y pequeña empresas estarán obligadas hacer uso de estas utilidades o ventajas si desean mantener o buscar un mayor acercamiento a sus posibles clientes el cual es pilar para desarrollarse exitosamente como entidades corporativas.

Por otro lado, es una realidad que la internet es un mar de datos donde cada día se están haciendo más denso, generandose millones de datos por segundo el cual se desarrollará la necesidad de optar por equipos, técnicas, y metodologías más eficientes ante un mundo donde la vanguardia es el valor que se puede hallar en los datos.

Referencias bibliográficas

Elmasri, R. & Shamkant N., (2002). *Fundamentos de Sistemas de Bases de Datos*. Conceptos fundamentales Tercera Edición, México.Addison Wesley.

Hernández, J., Ramírez, J & Ferri, C. (2005). *Introducción a la Minería de Datos*. España: Pearson.

https://msdn.microsoft.com/es-es/library/ms174949(v=sql.120).aspx

http://culturacrm.com/crm/recursos-crm/orange-data-mining-analisis-datos/

Anexos