Capítulo 2. Marco teórico del problema

2.1 Conceptos principales

Sistema de gestión de base de datos

Un sistema gestor de bases de datos (SGBD) consiste en una colección de datos interrelacionados y una colección de programas para acceder a esos datos. [8]

El objetivo principal de un SGBD es proporcionar un entorno que sea tanto conveniente como eficiente para las personas que lo usan para la recuperación y almacenamiento de la información. Una de las principales razones de usar SGBDs es tener un control centralizado tanto de los datos como de los programas que acceden a esos datos. [8]

Bases de datos relacionales

Es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer interconexiones (relaciones) entre los datos (que están guardados en tablas), y a través de dichas conexiones relacionar los datos de ambas tablas, de ahí proviene su nombre: "Modelo Relacional".

Aunque las bases de datos relacionales son la fuente de datos para la mayoría de aplicaciones de minería de datos, las técnicas de minería no son capaces de trabajar con toda la base de datos ó almacén de datos, sino que sólo son capaces de tratar con una sola tabla ó vista minable. [9]

Modelo entidad - relación

El modelo de datos entidad-relación (E-R) está basado en una percepción del mundo real que consta de una colección de objetos básicos, llamados entidades, y de relaciones entre estos objetos. Una entidad es una «cosa» u «objeto» en el mundo real que es distinguible de otros objetos. [8]

Sus principales características son:

Entidades: Cualquier información referente a un objeto que queremos almacenar
Atributos: Características del objeto que se quiere almacenar
Relaciones: Como interactúan unos con otros
Cardinalidad: Relación entre entidades

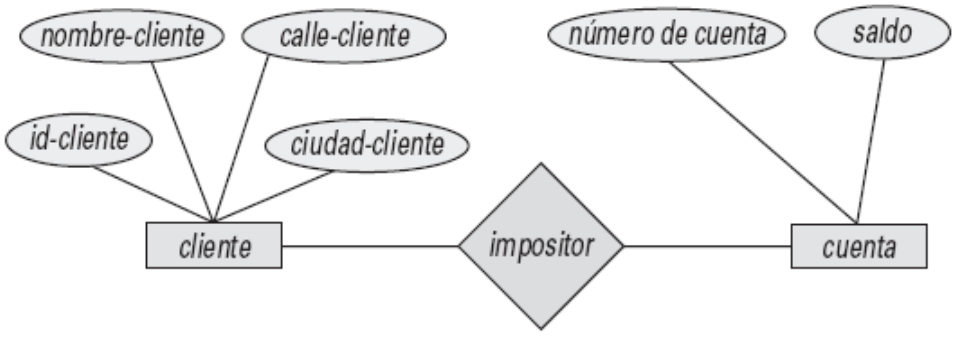

Figura 2.1.a Ejemplo del Modelo E-R [8]

Base de datos no volátil

La información no se modifica ni se elimina una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas. [10]

Base de datos intratable

Una base de datos intratable es aquella en la que las fases de análisis, diseño y desarrollo de la misma no han sido realizadas por un profesional de sistemas que conozca los métodos y procedimientos de cada una de dichas fases, haciendo que el tratamiento y consulta de los datos sea intratable. [10]

Almacenes de datos

Conjunto de datos históricos, internos o externos y descriptivos de un área de estudio, que están:

-integrados. [9]
-organizados. [9]
-variante en el tiempo. [9]

Para aplicar herramientas para:

-resumir [9]
-describir. [9]
-analizar los datos. [9]

Con el fin de ayudar en la toma de decisiones estratégicas.

Las ventajas fundamentales de un almacén de datos son:

- Su diseño específico. [9]

- Su separación de la base de datos transaccional. [9]

Diccionario de datos

- Descripción externa conceptual e interna de la base de datos. [11]

- Descripción de entidades, atributos y entidades externas. [11]

- Sinónimos de los datos contenidos en la base de datos. [11]

- Códigos de autorización y perfiles de los usuarios, así como los privilegios sobre la base de datos. [11]

Lenguaje de definición de datos

Un lenguaje de definición de datos (DDL) es un lenguaje proporcionado por el sistema de gestión de base de datos que permite a los usuarios de la misma llevar a cabo las tareas de definición de las estructuras que almacenarán los datos así como de los procedimientos o funciones que permitan consultarlos. [8]

Lenguaje de manipulación de datos

Un lenguaje de manipulación de datos (DML) es un lenguaje que permite a los usuarios acceder o manipular los datos organizados mediante el modelo de datos apropiado. [8]

Hay dos tipos básicamente:

• DMLs procedimentales. Requieren que el usuario especifique qué datos se necesitan y cómo obtener esos datos. [8]

• DMLs declarativos (también conocidos como DMLs no procedimentales). Requieren que el usuario especifique qué datos se necesitan sin especificar cómo obtener esos datos. [8]

Son DMLs: Select, Insert, Delete y Update [8]

Lenguaje de control de datos

Un Lenguaje de Control de Datos (DCL) es un lenguaje proporcionado por el Sistema de Gestión de Base de Datos que incluye una serie de comandos SQL que permiten al administrador controlar los privilegios en la Base de Datos. [11]

Lenguaje de control de transacciones

El Lenguaje de Control de Transacciones (Transaction Control Language – *TCL*) se utiliza para administrar los procesos transaccionales en una base de datos en relación a los requerimientos de atomicidad, consistencia, aislamiento y durabilidad, de igual forma permite regresar a su estado original cualquier transacción si esta no se lleva a cabo. [11]

Minería de datos

La minería de datos es 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 modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

De esta manera, dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos) y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado está íntimamente relacionada con la comprensibilidad del modelo inferido.

De una manera simplista pero ambiciosa, se puede decir que el objetivo de la minería de datos es convertir datos en conocimiento. [9]

Modelo de minería de datos

Patrones y tendencias que existen en los datos. Los modelos de minería de datos se pueden aplicar a situaciones empresariales como las siguientes:

- Predecir ventas. [12]
- Dirigir correo a clientes específicos. [12]
- Determinar los productos que se pueden vender juntos. [12]
- Buscar secuencias en el orden en que los clientes agregan productos a una cesta de compra. [12]

Sistema de minería de datos

Es una tecnología de soporte para el usuario final cuyo objetivo es extraer conocimiento útil y utilizable a partir de la información contenida en las bases de datos de las empresas. [9]

Vista SQL

Relaciones que no forman parte del modelo lógico pero se hacen visibles a los usuarios como relaciones virtuales. Se puede trabajar con gran número de vistas sobre cualquier conjunto dado de relaciones reales. [8]

Vista Minable

Es la combinación en una sola tabla de la información de varias tablas que requiramos para cada tarea concreta de minería de datos. [9]

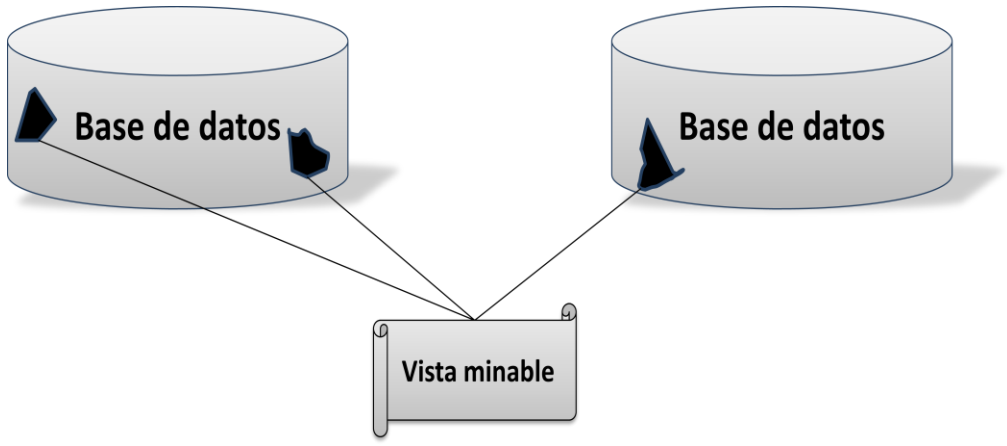

FIGURA 2.1.b Vista Minable

Herramientas de minería de datos

Las herramientas de minería de datos sirven para predecir tendencias y comportamientos. De esta manera permiten a las organizaciones tomar decisiones proactivas para adaptarse rápidamente a los cambios del mercado obteniendo así ventajas competitivas.

Las herramientas de minería de datos pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas por consultas en un sistema tradicional de soporte operacional. La potencialidad de estas herramientas reside en la capacidad de explorar las bases de datos en busca de patrones ocultos, encontrando información predecible que para un experto sería casi imposible debido al gran volumen de información.

Algunas herramientas de minería de datos son:

Matlab: Es un poderoso entorno de cálculo técnico integrado que combina el cálculo numérico, gráficos avanzados y visualización, y un lenguaje de programación de alto nivel. [13]

Weka: Es una extensa colección de algoritmos de máquinas de conocimiento desarrollados por la Universidad de Waikato (Nueva Zelanda) implementados en Java; útiles para ser aplicados sobre datos mediante las interfaces que ofrece. Weka ofrece también las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización. [14]

Aprendizaje automático

Área de inteligencia artificial que se ocupa de desarrollar algoritmos (programas) capaces de aprender, y constituye, junto con la estadística, el corazón del análisis inteligente de los datos.
Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema. [9]

Sistemas de aprendizaje automático

Tienen la intención de construir un modelo a partir de datos de entrada y cambiar su comportamiento de tal manera que son capaces de clasificar nuevos datos y desarrollarse mejor que en antiguas situaciones. [15]

Técnicas de minería de datos

Son el resultado de un largo proceso de investigación y desarrollo de productos orientados al almacenamiento, extracción y análisis de datos. [9]

Tipos de Datos

La minería de datos puede aplicarse a cualquier tipo de información, siendo las técnicas de minería diferentes para cada una de ellas. [9]

FIGURA 2.1.c Tipos de datos

Tipos de modelos producidos por la minería de datos

Modelos Predictivos:

Son aquellos que pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos variables objetivo o dependientes (target), usando otras variables o campos de la base de datos, a las que nos referiremos como variables independientes o predictivas. [9]

Modelos Descriptivos:

Son aquellos que pretender identificar patrones que explican o resumen los datos, además sirven para explorar las propiedades de los datos. [9]

Los sistemas para la toma de decisión

Son las diversas herramientas y sistemas que asisten a los directivos en la resolución de problemas y en la toma de decisiones. Su objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico (por ejemplo en medicina). [9]

Visualización de datos

Son técnicas de visualización que permiten al usuario:

- Descubrir.
- Intuir.
- Entender patrones.

Que serían más difíciles de "ver" a partir de descripciones matemáticas o textuales de los resultados. [9]

Fases de Descubrimiento de conocimiento en bases de datos (*Knowledge Discovery from Databases* - KDD)

Se define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- Válido: hace referencia a que los patrones deben seguir siendo válidos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención. [9]
- Novedoso: que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario. [9]
- Potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario. [9]
- Comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad). [9]

Como se deduce de la anterior definición, el KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos. [9]

A continuación se muestra un diagrama, el cual muestra las diferentes fases del KDD

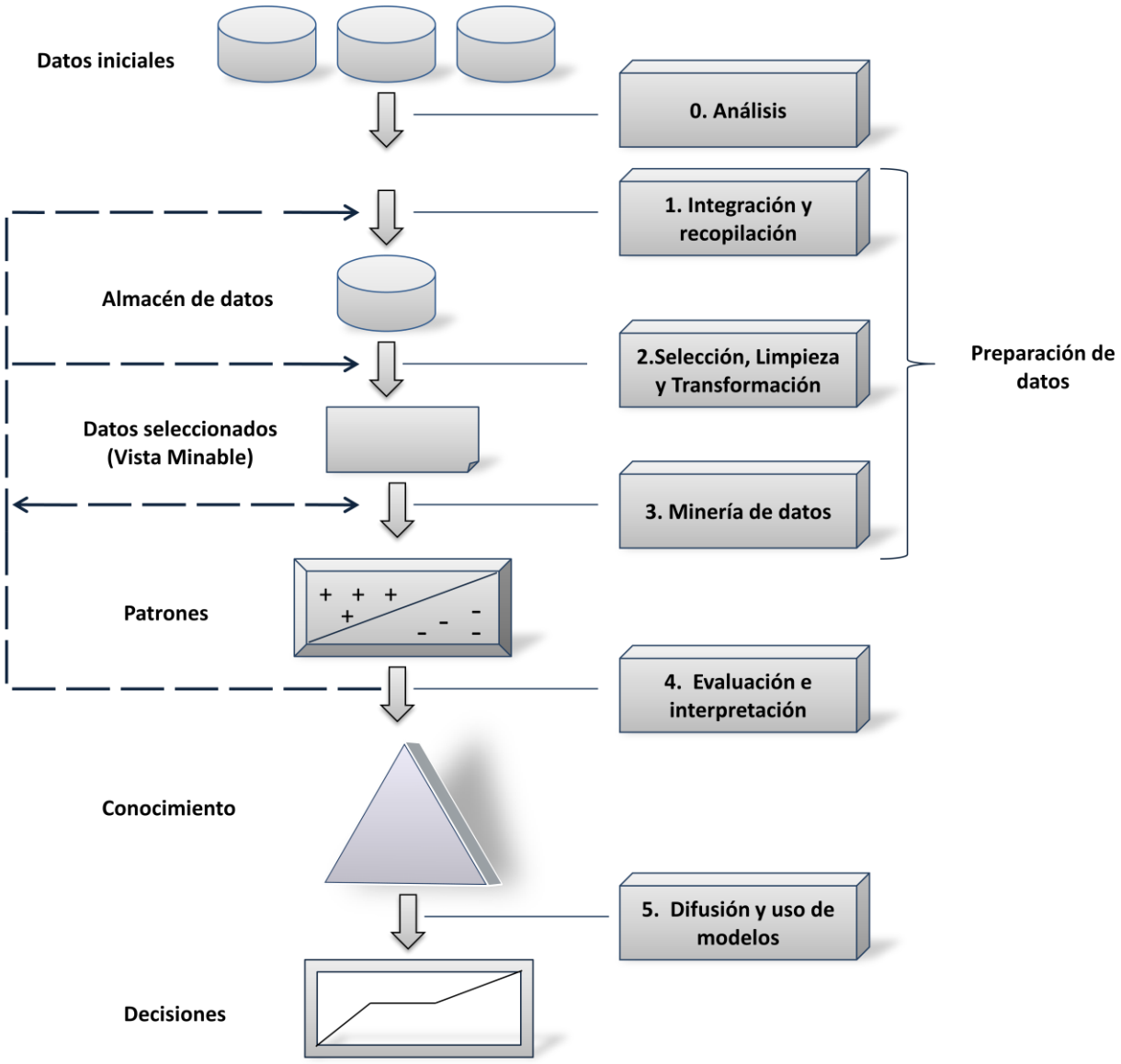

FIGURA 2.1.d KDD [9]

Así, los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso. [9]

Este proceso propone no sólo la definición de los problemas, sino también la obtención de los modelos, la evaluación y posible interpretación de los mismos. Sin embargo todo lo anterior no se muestra explícitamente, de manera que a continuación se muestra un esquema que se seguirá para cada uno de los análisis principales, cabe mencionar que dicho esquema parte después del segundo bloque y de igual forma que el primero no volverán a ser requeridos, debido a las limitantes mismas de este trabajo :

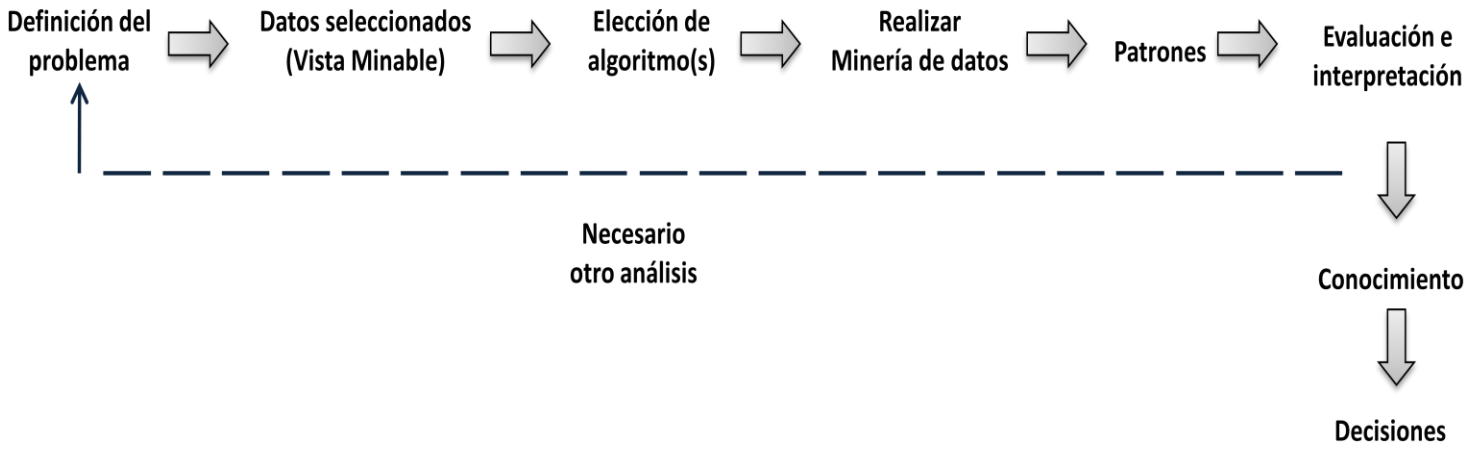

FIGURA 2.1.e Esquema para análisis

Definición del problema: Enunciado que describe la situación a ser tratada.

Datos seleccionados (Vista minable): Descripción (así como código) de la consulta que se utilizará para obtener los datos necesarios para la resolución del problema.

Elección del algoritmo (s): Exposición de razones para utilizar el o los algoritmos que brindaran los patrones a analizar posteriormente.

Minería de datos: Proceso de minería que consiste en la creación de la vista de origen de datos, selección de algoritmo (el cual se eligió en el punto anterior), características del mismo y descripción de los datos a procesar.

Patrones: Graficas, reglas y/o tablas con los resultados de la minería de datos.

Evaluación e interpretación: Define la precisión del análisis realizado utilizando los datos destinados a las pruebas, así mismo define si el análisis es suficiente para obtener el conocimiento necesario para resolver el problema, en caso de no existir patrones significativos es forzoso realizar otro análisis, cambiando los datos de entrada y/o salida o bien cambiando el algoritmo empleado.

Conocimiento: Descripción de la información obtenida durante todo el análisis. En este punto ya no es necesario ningún nuevo análisis pero tampoco se toman decisiones.

Decisiones: Finalización del análisis, en este punto se toman las decisiones basadas en el conocimiento adquirido en el apartado anterior, y las cuales buscan dar una solución al problema inicialmente planteado.

El esquema anterior puede ser asemejado al proceso de 5 pasos [17] para la toma de decisiones y resolución de problemas donde:

1er Paso. Definir el problema = Definición del problema y datos seleccionados (Vista Minable)
2º Paso. Buscar alternativas = Elección de algoritmos y realizar minería de datos
3er Paso. Valorar las consecuencias de cada alternativa = patrones, evaluación e interpretación
4º Paso. Elegir la mejor alternativa = Conocimiento
5º Paso. Aplicar alternativa escogida = Decisiones

Media

Denominada más frecuentemente como "promedio". Es la suma de los valores obtenidos entre la cuenta de observaciones. [16]

Mediana

Es el valor medio de un conjunto de números. La mitad de los números tiene un mayor valor que la mediana, y la otra mitad tiene n valor inferior a la mediana. [16]

Desviación estándar

Medida de la dispersión de los datos. De manera específica, las medidas de desviación estándar miden la manera en que se extienden los valores de datos a partir de la media. [16]

Rango

La diferencia entre el valor más pequeño y es más grande de un valor de datos. [16]

Distribución

La serie de valores observada para un atributo determinado en sus datos. La distribución suele mostrarse de manera útil utilizando histogramas o gráficas circulares. [16]

Datos continuos

Datos cuyo espacio muestral especificado como dominio no es numerable. Por ejemplo, todos envejecemos en fracciones de segundo. [16]

Datos discretos

Datos cuyo espacio muestral especificado como dominio es numerable. Algunos ejemplos son país de residencia o estado civil. [16]

Algoritmo

Método para resolver un problema, debe presentarse como una secuencia ordenada de instrucciones que siempre se ejecutan en un tiempo finito y con una cantidad de esfuerzo también finito. En un algoritmo siempre debe haber un punto de inicio y un punto de terminación, estos deben ser únicos y deben ser fácilmente identificables. [9]

Todo algoritmo debe cumplir las siguientes características:

A. Debe ser Preciso; esto es, debe especificar sin ambigüedad el orden en que se deben ejecutar las instrucciones.

B. Debe estar Definido; esto es, cada vez que se ejecute bajo las mismas condiciones, la secuencia de ejecución deberá ser la misma proporcionándonos el mismo resultado.

C. Debe ser Finito; esto es, siempre que sea adecuado se realizarán un número finito de instrucciones, en un tiempo finito y requiriendo una cantidad finita de esfuerzo.

Proceso incremental

Proceso por el cual un proyecto es realizado por partes y que al final terminará siendo la solución completa requerida para un problema. [9]

Procesos Iterativos

Proceso en el que un sistema mejora su funcionalidad durante cada iteración. [9]

2.2 Definición de tareas de minería

Tareas y métodos

Una tarea de minería de datos es un (tipo de) problema de minería de datos. Por ejemplo, "clasificar las piezas del proveedor Minatronix en óptimas, defectuosas reparables y defectuosas irreparables" es una tarea de clasificación, que podría resolverse mediante árboles de decisión o redes neuronales, entre otros métodos. Es muy importante distinguir el problema de los métodos para solucionarlo. Una tarea puede tener muchos métodos diferentes para resolverla y el mismo método puede resolver muchas tareas. [9]

Tareas predictivas

Se trata de problemas en los que hay que predecir uno o más valores para uno o más ejemplos. [9]

Tareas descriptivas

Los ejemplos se presentan como un conjunto sin etiquetar ni ordenar de ninguna manera, el objetivo, es describir los existentes. [9]

Se pueden considerar tareas descriptivas:

A) Las tablas de frecuencias.

B) El análisis de componentes principales) .

C) Agrupamiento (clustering).

Su objetivo es obtener grupos o conjuntos entre los elementos, de tal manera que los elementos asignados al mismo grupo sean similares.

FIGURA 2.2.a Tareas de minería

Tareas de clasificación

La clasificación (predictiva) es quizá la tarea más utilizada.

- Cada instancia (o registro de la base de datos) pertenece a una clase, la cual se indica mediante el valor de un atributo que llamamos la clase de la instancia.
- El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase.

- El objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase.

Ejemplo: Un oftalmólogo desea clasificar nuevos pacientes, para decidir si es conveniente operarlos o no en función de una base de datos de sus antiguos pacientes. [9]

Tareas de regresión

La regresión (predictiva), la principal diferencia respecto a la clasificación es que el valor a predecir es numérico.

- Su objetivo es minimizar el error entre el valor predicho y el valor real.

Ejemplo: Un empresario quiere conocer cuál es el costo de un nuevo contrato basándose en los datos correspondientes a contratos anteriores. [9]

Tareas de agrupamiento (segmentación)

El agrupamiento (clustering) es descriptiva, consiste en obtener grupos "naturales" a partir de los datos. Se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo. [9]

Ejemplo, una librería identifica grupos de clientes en base a sus preferencias, para que recomiende otros libros comprados por clientes de su mismo grupo. [9]

Tareas de asociación

Las reglas de asociación (descriptiva) son similares a las de las correlaciones, tienen como objetivo identificar relaciones no explícitas entre atributos categóricos.

Estas reglas pueden ser de muchas formas, aunque la formulación más común es del estilo

"Si el atributo X toma el valor d entonces el atributo Y toma el valor b".

Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados.

Por ejemplo en el análisis de la cesta de la compra, para identificar productos que son frecuentemente comprados juntos, información que puede usarse:

- Para ajustar los inventarios.

- Para la organización física del almacén.

- En campañas publicitarias.

Ejemplo: En una tienda de electrodomésticos, se analizan ventas y se descubre que el 30% de los clientes que compraron un televisor hace 6 meses compraron un DVD en los siguientes 2 meses. [9]

Tareas de correlación

Las correlaciones (descriptiva) se usan para examinar el grado de similitud de los valores de dos variables numéricas, para medir la correlación lineal es con el coeficiente de correlación r, -1 y 1. [9]

- Si es 0 no hay correlación.

- Cuando r es positivo, las variables tienen un comportamiento similar (ambas crecen o decrecen al mismo tiempo).

- Cuando r es negativo si una variable crece la otra decrece.

Ejemplo: Un inspector de incendios obtiene correlaciones negativas entre el empleo de aisladores y la frecuencia de incendios. [9]

Algoritmos (técnicas) de minería de datos

El algoritmo de minería de datos es el mecanismo que crea modelos de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos, buscando patrones y tendencias específicos. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos. [19]

El modelo de minería de datos que crea un algoritmo puede tomar diversas formas, incluyendo:

- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción.

- Un árbol de decisión que predice si un cliente determinado comprará un producto.

- Un modelo matemático que predice las ventas.

- Un conjunto de clústeres que describe cómo se relacionan los escenarios de un conjunto de datos.

La elección del algoritmo apropiado para una tarea específica puede ser un trabajo difícil. Aunque se pueden utilizar diferentes algoritmos para realizar una misma tarea, cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un tipo de resultado.

Tampoco es necesario que los algoritmos sean usados de modo independiente: en una solución de minería de datos se pueden utilizar algunos algoritmos para examinar los datos y, después, usar otros para predecir un resultado específico basándose en esos datos. Por ejemplo, se puede utilizar un algoritmo de clústeres, que reconoce patrones, para dividir los datos en grupos que sean más o menos homogéneos, y luego usar los resultados para crear un mejor modelo de árbol de decisión.

La siguiente figura muestra algunas técnicas de minería de datos así como las tareas en las que pueden ser utilizadas para obtener los modelos predictivos y descriptivos.

NOMBRE DE TÉCNICA	TAREAS PREDICTIVAS		TAREAS DESCRPTIVAS		
	Clasificación	Regresión	Agrupamiento	Reglas de Asociación	Correlaciones
Redes Neuronales	X	X	X		
Puntuación de grado de interés	X				
Bayesiano con prioridad K2	X				
Regresión lineal y logarítmica		X			X
Kmeans			X		
EM			X		
Apriori			X	X	
Naive Bayes	X				

FIGURA 2.2.b Técnicas y tareas de minería [9]

2.3 Hipótesis

Como principales hipótesis para el presente trabajo de investigación podemos encontrar que:

- La tutoría alcanza a un 70% de la población de alumnos de licenciatura en ingeniería.
- Aquellos alumnos que no obtuvieron una buena calificación en su examen diagnóstico son aquellos que asisten con mayor regularidad a la tutoría.
- La tutoría influye de manera considerable en el promedio del alumno.
- La distancia hogar-universidad es uno de los principales factores en la inasistencia del alumno a la tutoría, entre mas distancia mayor inasistencia.
- La asistencia del alumno a la tutoría va ligada directamente a la disponibilidad del alumno, esto es aquellos que trabajan tienen una asistencia menor que aquellos que no lo hacen.
- La clase social de un alumno no influye en la asistencia del alumno en la tutoría ni su promedio del semestre.
- El desempeño de un alumno va de la mano con la efectividad que tenga un tutor en las sesiones de tutoría.
- Los hábitos de estudio de un alumno son parte fundamental en un buen desempeño académico.