MINERIA DE DATOS

inteligencia de negocios hace referencia a un conjunto de productos y servicios para acceder a los datos para acceder a los datos analizarlos y convertirlos en información a partir de datos mediante herramientas de extracción consulta y herramientas de modelo de datos es un proceso de extracción automatizada de información sintetizada relevante y con capacidad predictiva a partir de grandes bases de datos abarca diversas técnicas enfocadas a la la visualización análisis y modelización de procesos a partir de grandes cantidades de datos. Su objetivo es identifica patrones significativos que sean validos, novedosos y potencialmente útiles para el negocio entendibles aplicando técnicas de inteligencia artificial desde el punto de vista disciplinar este proceso puede definirse como multidisciplinar ya que abarca conceptos y prácticas de estadística aplicando inteligencia artificial sobre repositorios de datos específicamente bases de datos multidimensionales.

Un proceso típico de minería de datos entendiéndose como tal un conjunto de etapas sucesivas consta de los siguientes pasos generales: a partir de una base de datos transaccional se ejecuta primero una selección de un conjunto de datos que incluyan las variables que se quiere predecir calcular o inferir y las que sirven para hacer el cálculo o proceso luego se preparan los datos para detectar la presencia de valores atípicos y la ausencia de datos o valores nulos se procede a la transformación del conjunto de datos de entrada también se la conoce como pre procesamiento de los datos ya limpios y válidos para almacenarlos en data Waterhouse. La la selección y aplicación de la técnica de minería de datos construye el modelo predictivo o descriptivo seleccionando una técnica de minería apropiada y finalmente el quinto extracción de conocimiento interpretación y evaluación de datos como resultado de la técnica aplicada se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables una vez obtenido el modelo se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias en el caso de haber obtenido varios modelos mediante el uso de distintas técnicas se deben comparar los modelos en busca de aquel que se ajuste mejor al problema si ninguno de los modelos alcanzan los resultados esperados debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Tenemos ahora la diferencia que hay entre consultas que se hagan a la base transaccional y la minería de datos en los sistemas estándar de gestión de bases de datos transaccionales las consultas se resuelven accediendo a distintos conjuntos de datos almacenados por ejemplo ventas del último mes de un producto el usuario empieza con una consulta o una suposición explora los datos y construye un modelo los sistemas de data mining infieren conocimiento de la base de datos en forma de estructuras y patrones por ejemplo qué clientes son potenciales compradores de un producto el usuario no propone modelos el sistema los encuentra y lo sugiere al usuario.

Esquema de una arquitectura típica del proceso de minería de datos en la fuente de datos es el repositorio de los datos ya transformados como lo son los Data Warehouse multidimensionales el servidor de la base de datos se utiliza en los sistemas en red para gestiornar la informacion la base de conocimiento guía la búsqueda evaluando los patrones metadatos y la obtención de conocimiento previo de los datos. Surgieron a partir de la investigación en como respuesta a las necesidades canciones de esta disciplina planteaban contienen elementos de conocimiento normalmente en la forma de hechos y reglas de inferencia así como la manera en que estos han de ser utilizado el algoritmo de minería de datos permite realizar diferentes tipos de análisis como caracterización asociación clasificación análisis de grupos evolución y análisis de desviaciones los veremos luego la evaluación de patrones son medidas para conocer la relevancia de los patrones de no ser significativa sé ejecutar el proceso de minería de datos enfocando en variables diferentes para obtener un nuevo modelo que pueda ser relevante la interfaz gráfica finalmente permite la interacción con el usuario y la visualización de los patrones hallados bajo distintas formas.  
Marcas de minería de datos modelos que pueden ser predictivos o descriptivos. Un modelo predictivo responde preguntas sobre un comportamiento futuro de los datos un modelo descriptivo en cambio proporciona información sobre las relaciones encontradas entre los datos y sus características para poder entenderlos mejor.  
En base a una clasificación se evalúa mediante modelos correlacionales las probabilidades de un resultado Por ejemplo si de acuerdo a determinados atributos o el tipo de dolencia que puede tener un paciente de acuerdo a los síntomas en base a una regresión lineal Qué es la más utilizada se encuentran relaciones entre variables por ejemplo calcular el tiempo previsible que se empleará en corregir los errores de un desarrollo de software los modelos estadísticos son una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales para indicar las diferentes factores que modifican la variable de respuesta por ejemplo análisis de varianza discriminante dispersión etc la secuenciación intenta predecir el en función de su evolución en el tiempo a partir de este conocimiento y bajo el supuesto de que no van a producirse cambios estructurales por ejemplo estimar la demanda de energía eléctrica para un periodo determinado.  
Enseñando con las técnicas predictivas tenemos el árbol de decisión sobre los datos se va realizando sucesivas bifurcaciones hasta llegar a un resultado modelo discriminante de entre Los criterios seleccionados aquella que permite a los distintos distintos grupos del criterio base se obtiene de este modo la primera segmentación y se realizan nuevas segmentaciones de cada uno de los segmentos resultantes en forma iterativa pasajeros. pasajeros del Titanic tuvieron más probabilidades de sobrevivir a su hundimiento y Qué características tuvieron asociadas a la supervivencia del naufragio un procedimiento de árbol de decisión selecciona automáticamente los grupos homogéneos con la mayor diferencia en proporción de supervivientes en esta caso ha encontrado el sexo: hombres y mujeres.

El siguiente paso subdivide cada uno de estos grupos en función de otra característica resultando que los hombres son divididos en adultos y niños, mientras que las mujeres se dividen en grupos basados en la calse en la que viaban en el barco. Si un pasajero del titanic es hombre y es adulto tiene una probabilidad de sobrevivir según este modelo del 20% la probabilidad de supervivencia de cada uno de las subdivisiones puede utilizarse con fines predictivos para vaticinar el grado de probabilidad de supervivencia de los miembros de cada grupo.

las reglas de asociación encuentran interrelaciones entre datos generando reglas del tipo que pasa así a partir de un proceso de aprendizaje de modelo por eje ejemplo conocer en un hipermercado si un cliente que compra leche muy probablemente compre también pan se utilizan los llamados algoritmos de aprendizaje supervisados que se basan en entrenar un modelo por medio de diferentes datos conocidos para poder predecir un dato desconocido a priori trabaja sobre 2 tablas la de aprendizaje o entrenamiento y la tabla de verificación en donde se aplica el modelo generado a partir de la tabla de aprendizaje para ver la calidad de este modelo para una correcta predicción

las redes neuronales son un paradigma de aprendizaje y procesamiento automatizado inspirado en la forma en que funciona el sistema nervioso humano eh cada centro neuronal recibe múltiples estímulos simultáneos que determinan una única salida se trata de un sistema de interconexión de nodos neuronales en una red que colabora para producir un único estímulo de salida se estructuran como un grupo numeroso de procesadores conectados entre sí y trabajando en paralelo aprende a través del entrenamiento las tareas básicas de las redes neuronales son reconocer clasificar agrupar asociar almacenar patrones y someter a los datos a un tratamiento mediante algoritmos que dan lugar a que se obtengan resultados de índole predictiva.

En cuanto a las técnicas usadas por los modelos descriptivos tenemos el agrupamiento o clustering que clasifica elementos en grupos definidos por el usuario elementos en grupos definidos por el usuario en base Asus características agrupando los datos según indicadores de frecuencia en segmentos de comportamiento homogéneo el objetivo es maximizar la similitud de elementos de plaster y las diferencias entre clusters cada elemento debe estar asignado a un único clúster su finalidad esencial es revelar concentraciones en los datos para su agrupamiento eficiente en estos clusters según su homogeneidad pudiendo utilizarse variables cualitativas o cuantitativas el agrupamiento se realiza basándose en la proximidad o lejanía de unos con otros es esencial el uso adecuado del concepto de distancia por ejemplo en los análisis de clientes con tarjeta de crédito de tipo gold eh por género hijos estado civil e ingresos dio como resultado este ejemplo de agrupamiento.

Los métodos simétricos o de aprendizaje no supervisado trabaja sobre elementos clasificados inductivamente en grupos que no son conocidos con anterioridad a diferencia de la técnica de clustering donde son definidos a priori por el usuario de modo que todas las variables son tratadas en el mismo nivel de importancia y no hay una hipótesis previa de causalidad para proseguir