“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA”

UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

FACULTAD DE INGENIERÍA

ESCUELA DE FORMACIÓN PROFESIONAL DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

MINERÍA DE DATOS

“Introducción a la Minería de Datos y el Conocimiento de los Datos: Un Análisis del Capítulo I de Data Mining: Concepts and Techniques de Han, Kamber y Pei**”**

Autor (res):

CORDOVA CONDOR, Harisa Meylin

OLAZO DIAZ, Fiorella Aracely

POMA PALACIOS, Alisson Naily

TORRES LUCAS, Luis Josue

Docente:

Mg. MUÑOZ ROBLES, Williams Antonio

Cerro de Pasco - Perú - 2025

ÍNDICE:

INTRODUCCIÓN

En las últimas décadas, la humanidad ha sido testigo de un crecimiento exponencial en la generación de datos. Empresas, instituciones científicas, organizaciones gubernamentales y usuarios comunes producen diariamente enormes volúmenes de información, provenientes de transacciones comerciales, redes sociales, investigaciones médicas, sensores tecnológicos y múltiples actividades de la vida cotidiana. Este fenómeno ha dado lugar a lo que hoy conocemos como la era de los datos, en la cual la disponibilidad masiva de información plantea tanto oportunidades como desafíos.

Sin embargo, disponer de grandes cantidades de datos no es suficiente si no se cuenta con las herramientas y técnicas adecuadas para analizarlos y convertirlos en conocimiento útil. Frente a esta necesidad surge la minería de datos (data mining), también denominada descubrimiento de conocimiento a partir de datos (KDD, por sus siglas en inglés). Esta disciplina, relativamente joven pero en rápido crecimiento, combina enfoques de la informática, la estadística, la inteligencia artificial y otras áreas afines para extraer patrones significativos, relaciones ocultas y modelos predictivos a partir de conjuntos de datos complejos.

El libro Data Mining: Concepts and Techniques de Jiawei Han, Micheline Kamber y Jian Pei, considerado una de las referencias fundamentales en el área, dedica su primer capítulo a establecer los conceptos básicos de esta disciplina. Allí se explican no solo las definiciones y alcances de la minería de datos, sino también su relación con el proceso integral de descubrimiento de conocimiento, los tipos de datos que pueden analizarse, las diferentes formas de conocimiento que es posible extraer, así como sus principales aplicaciones e implicancias sociales.

En particular, los apartados iniciales de este capítulo permiten comprender que la

minería de datos es más que una simple técnica: constituye un proceso estratégico para transformar datos masivos en información organizada y, finalmente, en conocimiento que apoye la toma de decisiones en diversos campos. Casos como el proyecto Google Flu Trends, que utilizó patrones de búsqueda en la web para anticipar brotes gripales, ilustran cómo esta disciplina puede tener impactos directos en la ciencia, la salud y la sociedad en general.

El presente trabajo monográfico tiene como propósito analizar y sistematizar los principales conceptos del Capítulo I del libro de Han, Kamber y Pei, con el fin de ofrecer una visión integral de la minería de datos como disciplina clave en el contexto de la sociedad digital. Para ello, se abordará qué es la minería de datos y por qué resulta necesaria, se explicará su papel dentro del proceso de descubrimiento de conocimiento, y se revisarán sus aplicaciones más relevantes. Finalmente, se reflexionará sobre su importancia para enfrentar los retos de la era actual, en la cual la capacidad de transformar datos en conocimiento se ha convertido en un factor esencial para la innovación y el desarrollo.

DESARROLLO:

1. ¿Qué es la minería de datos?

1.1. Definición general.

La minería de datos es el proceso de descubrir patrones, modelos y conocimiento en grandes volúmenes de datos mediante métodos inteligentes y técnicas analíticas avanzadas. Su finalidad es transformar grandes cantidades de datos aparentemente desorganizados en información estructurada y conocimiento útil que facilite la toma de decisiones. El término data mining se popularizó en la década de 1990 como una metáfora de "buscar pepitas de oro en los datos", reflejando la idea de extraer elementos valiosos de un conjunto masivo de información.

1.2. Importancia en la era de los datos.

En la actualidad vivimos en la llamada era de los datos, caracterizada por la explosión de información generada por las redes sociales, los negocios electrónicos, los sensores inteligentes, la investigación científica, la salud y prácticamente todas las áreas de la sociedad moderna. Esta abundancia de datos plantea un gran reto: no basta con almacenarlos, sino que es necesario analizarlos, interpretarlos y convertirlos en conocimiento aplicable.

En este contexto, la minería de datos se ha convertido en una herramienta estratégica, ya que permite identificar tendencias ocultas, predecir comportamientos, optimizar procesos y generar valor en múltiples sectores.

1.3. Ejemplo práctico (como Google Flu Trends).

- **Contexto:** Cada día, millones de personas buscan información en Google. Esos datos no solo sirven para mostrar resultados, sino que esconden patrones colectivos.
- **Descubrimiento:** Google detectó que cuando muchas personas buscaban palabras relacionadas con "síntomas de gripe", coincidía con un aumento real de enfermos en hospitales.
- **Aplicación:** Con ese patrón, crearon Google Flu Trends, una herramienta que podía predecir la propagación de la gripe hasta 2 semanas antes que los métodos tradicionales de salud pública.
- **Conclusión:** La minería de datos convierte simples consultas en información útil y anticipada para resolver problemas de gran escala (salud, seguridad, tendencias sociales, etc.).

1.4. Diferentes términos usados: data mining, KDD, descubrimiento de patrones, etc.

La minería de datos se ha descrito con distintos nombres y enfoques a lo largo del tiempo, entre ellos:

- **Data Mining:** término más difundido desde los años noventa.
- **KDD (Knowledge Discovery from Data):** descubrimiento de conocimiento a partir de datos, que abarca el proceso completo de limpieza, integración, transformación, minería, evaluación y presentación.

- **Descubrimiento de patrones:** destaca la identificación de tendencias y relaciones ocultas.
- **Extracción de conocimiento / Arqueología de datos:** resaltan la idea de explorar y "desenterrar" información valiosa.
- **Analítica de datos:** concepto moderno vinculado con inteligencia de negocios y análisis predictivo en entornos empresariales.

2. Minería de datos como parte del proceso de descubrimiento de conocimiento (KDD)

2.1. Explicar la diferencia entre minería de datos y KDD.

Aunque en muchos contextos los términos minería de datos y descubrimiento de conocimiento a partir de datos (KDD, por sus siglas en inglés) se utilizan como sinónimos, en sentido estricto no son equivalentes.

El KDD representa el proceso completo que convierte grandes volúmenes de datos en conocimiento útil y organizado.

La minería de datos, en cambio, es solo una de las etapas de ese proceso, la más relevante, ya que es en ella donde se aplican algoritmos inteligentes para extraer patrones, tendencias o modelos ocultos en los datos.

2.2. Describir el proceso completo (con sus fases: limpieza, integración, transformación, selección, minería, evaluación y presentación).

Vivimos en la era de los datos, más que en la de la información. Actualmente, se generan terabytes y hasta petabytes de datos diariamente, provenientes de diversas fuentes:

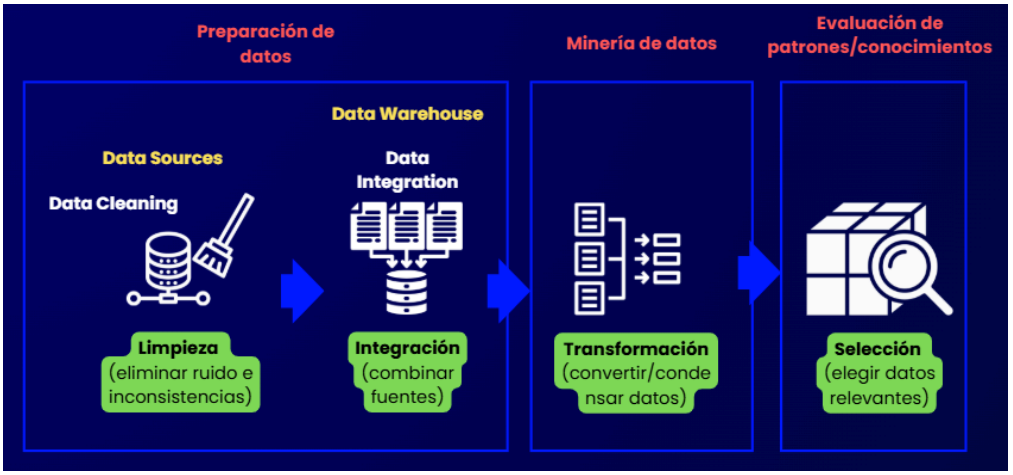
- Negocios: transacciones de ventas, perfiles de clientes, registros financieros.
- Ciencia e ingeniería: sensores, mediciones de procesos, experimentos, observaciones de sistemas.
- Salud: datos genómicos, historiales clínicos, monitoreo de pacientes, imágenes médicas.
- Internet: búsquedas en la web, redes sociales, fotografías, videos, comunidades virtuales.

Frente a este escenario, surge la necesidad de contar con herramientas capaces de transformar datos masivos en conocimiento organizado. El KDD aborda esta necesidad a través de un proceso iterativo y estructurado que incluye las siguientes fases:

1. **Limpieza de datos** → eliminar ruido, errores e inconsistencias.
2. **Integración de datos** → combinar información proveniente de múltiples fuentes.
3. **Transformación de datos** → convertir y consolidar los datos en formatos adecuados (resúmenes, agregaciones, normalizaciones).
4. **Selección de datos** → escoger solo la información relevante para el análisis.
5. **Minería de datos** → aplicar algoritmos y técnicas inteligentes para descubrir patrones y modelos.
6. **Evaluación de patrones/modelos** → **i**dentificar cuáles son realmente

interesantes o representativos.

7. **Presentación del conocimiento** → visualizar y comunicar los resultados de manera comprensible para los usuarios.



Este conjunto de fases permite que los datos en bruto pasen a convertirse en conocimiento organizado, que puede ser utilizado en la toma de decisiones estratégicas.

2.3. Señalar que muchas veces “data mining” se usa como sinónimo de KDD, aunque en teoría es solo una parte.

En la práctica, tanto en la industria como en la investigación y los medios, el término data mining suele emplearse para referirse a todo el proceso de descubrimiento de conocimiento. Esto se debe a que es más corto y popular que la denominación completa KDD. Sin embargo, es importante reconocer que, en sentido riguroso, la minería de datos es solo una parte del proceso, aunque se considere la etapa central y más decisiva para extraer valor de los datos masivos.

3. Diversidad de Tipos de Datos para la Minería de Datos

La minería de datos se aplica a cualquier tipo de datos siempre que sean significativos, pero cada tipo requiere métodos específicos.

Tipos de Datos:

- **Estructurados:** Datos con estructuras claras, como bases de datos relacionales o cubos de datos (ejemplo: tablas con columnas fijas como nombre, precio, etc.).
- **Semiestructurados:** Datos con cierta estructura, pero más flexibles, como transacciones (ejemplo: lista de compras en una tienda) o datos de grafos/redes (ejemplo: redes sociales donde los nodos son personas y las aristas son conexiones).
- **No estructurados:** Datos sin estructura definida, como texto, imágenes o videos (ejemplo: reseñas de productos o videos publicitarios).
- **Mezcla de datos:** En la vida real, los datos suelen combinar los tres tipos (ejemplo: un sitio web de compras con datos estructurados de productos, texto de reseñas y videos promocionales).

Ejemplo práctico: Una tienda en línea que analiza datos de compras (transacciones estructuradas), reseñas de clientes (texto no estructurado) y redes de recomendaciones (grafos semiestructurados).

4. Minería de Diversos Tipos de Conocimiento

La minería de datos puede descubrir diferentes tipos de conocimientos, divididos en tareas descriptivas (caracterizan datos) y predictivas (hacen predicciones).

4.1. Sumarización de datos multidimensionales

- **Qué es:** Resumir grandes conjuntos de datos en un formato comprensible, a menudo en espacios multidimensionales (ejemplo: por categoría de producto, ubicación o tiempo).
- **Cómo se hace:** Usando **cubos de datos** para agregar información, permitiendo a los usuarios explorar (drill-down o roll-up) con facilidad.
- **Ejemplo:** En una tienda como Rappi, podrían resumir ventas por categoría (comida, productos), ubicación (Lima, Arequipa) y tiempo (semana pasada), mostrando un gráfico de pastel con las ventas más altas.

4.2. Minería de patrones frecuentes, asociaciones y correlaciones

Los patrones frecuentes son combinaciones de ítems, secuencias o estructuras que aparecen a menudo en los datos.

- **Conjuntos de ítems frecuentes:** Ítems que se compran juntos (ejemplo: leche y pan).
- **Subsecuencias frecuentes:** Secuencias de acciones (ejemplo: comprar laptop, luego bolsa).
- **Subestructuras frecuentes:** Estructuras como grafos (ejemplo: conexiones en redes sociales).

Las reglas de asociación identifican relaciones entre ítems:

- **Ejemplo unidimensional:** "computadora $\Rightarrow$ webcam" (soporte = 1%, confianza = 50%).
- **Ejemplo multidimensional:** "edad 20-29 $\wedge$ ingresos 40K-49K $\Rightarrow$

compra laptop" (soporte = 0.5%, confianza = 60%).

Soporte: Porcentaje de transacciones con ambos ítems.

Confianza: Probabilidad de que un ítem se compre dado otro.

Ejemplo práctico:

En Rappi, podrían descubrir que el 2% de los pedidos incluyen arroz y pollo (soporte), y que el 60% de quienes piden arroz también piden pollo (confianza). Esto ayuda a recomendar productos.

4.3. Clasificación y regresión para análisis predictivo

Clasificación: Crear un modelo para asignar etiquetas categóricas a datos desconocidos usando datos de entrenamiento.

Formas de representar el modelo (ver Figura 1.2):

- **(a) Reglas SI-ENTONCES:** Ejemplo: "Si edad = joven Y ingreso = alto → clase A".
- **(b) Árbol de decisión:** Divide datos según atributos (edad, ingreso) para predecir clases.
- **(c) Red neuronal:** Usa conexiones ponderadas para clasificar.

Regresión: Predecir valores numéricos continuos (ejemplo: ingresos de un producto).

Ejemplo práctico:

En Yape, podrían clasificar usuarios en 'frecuentes' o 'esporádicos' según sus transacciones, o predecir cuánto gastará un usuario en un mes con regresión.

4.4. Análisis de clústeres

Clustering: Agrupar datos basándose en similitud, sin etiquetas previas. Maximiza la similitud dentro de un grupo y minimiza entre grupos.

Ejemplo: Identificar grupos de clientes en una tienda según patrones de compra.

Uso: Generar clases o taxonomías para marketing.

Ejemplo práctico:

En Rappi, el clustering puede identificar grupos como ‘clientes que piden comida rápida’ o ‘clientes que compran productos saludables’ para personalizar promociones.

4.5. Aprendizaje profundo

Usa redes neuronales para generar características automáticamente (ingeniería de características).

Ejemplo: Predecir brotes de enfermedades usando datos de salud (casos diarios, hospitalizaciones) para crear tasas semanales significativas.

Aplicaciones: Clasificación, clustering, detección de anomalías, visión por computadora, procesamiento de lenguaje natural.

Ejemplo práctico:

En una app como Yape, el aprendizaje profundo podría analizar patrones de transacciones para detectar fraudes o recomendar servicios personalizados.

4.6. Análisis de valores atípicos

- **Valores atípicos:** Datos que no siguen el comportamiento general (anomalías).
- **Ejemplo:** Compras inusualmente grandes en una tarjeta de crédito indican posible fraude.
- **Métodos:** Pruebas estadísticas, medidas de distancia o densidad.

Ejemplo práctico:

En un banco, un valor atípico podría ser una transferencia de S/10,000 desde una cuenta que normalmente mueve S/100, lo que podría indicar un fraude.

4.7. ¿Son todos los resultados de minería interesantes?/¿Qué hace que un patrón sea interesante?

Un patrón es interesante si es:

- Fácil de entender.
- Válido en datos nuevos.
- Útil y novedoso.
- Valida hipótesis o es accionable (ejemplo: “un gran terremoto sigue a pequeños sismos”).

Medidas objetivas:

Soporte:

soporte(X=>Y)=P(XUY)

- Porcentaje de transacciones con X e Y.
- **Ejemplo:** 1% de compras incluyen computadora y webcam.

Confianza:

confianza(X=>Y)=P(Y|X)

- Probabilidad de que Y ocurra si X está presente.
- **Ejemplo:** 50% de quienes compran computadora compran webcam.

Medidas subjetivas: Patrones relevantes para el usuario (ejemplo: útiles para un gerente de marketing, no para un analista de empleados).

Umbrales: Reglas con soporte/confianza bajos se descartan como ruido.

Desafíos:

- Generar todos los patrones es ineficiente.

- Solución: Usar restricciones para enfocarse en patrones interesantes.

Ejemplo práctico:

En Rappi, una regla como ‘arroz ⇒ pollo [soporte=2%, confianza=60%]’ es interesante si ayuda a recomendar productos, pero si el soporte es muy bajo (0.1%), podría ser solo un caso raro.

La minería de datos es versátil porque maneja diversos tipos de datos (estructurados, semiestructurados, no estructurados) y produce diferentes conocimientos (sumarización, patrones, predicciones, clústeres, anomalías).

5. Minería de datos:confluencia en múltiples disciplinas

La minería de datos es una disciplina que estudia métodos eficientes y efectivos para descubrir patrones y obtener conocimiento a partir de grandes volúmenes de datos en diversas aplicaciones.

Por esta razón, se considera un punto de encuentro de varias disciplinas, entre ellas:

- Aprendizaje automático (machine learning)
- Estadística
- Reconocimiento de patrones
- Procesamiento de lenguaje natural (PLN)
- Tecnología de bases de datos
- Visualización de datos e interacción humano-computadora (HCI)
- Algoritmos y computación de alto rendimiento
- Ciencias sociales
- Diversos dominios de aplicación

La naturaleza interdisciplinaria de la minería de datos ha sido clave para su éxito y sus amplias aplicaciones.

Al mismo tiempo, la minería de datos no solo se nutre de estas disciplinas, sino que también, gracias a sus avances en el análisis de grandes volúmenes de datos, ha llegado a influir y aportar al desarrollo de estas mismas áreas en los últimos años.

5.1. Estadística y minería de datos

La estadística se ocupa de la recolección, análisis, interpretación y presentación de datos. La minería de datos está estrechamente relacionada con esta disciplina.

Un modelo estadístico es un conjunto de funciones matemáticas que describen el comportamiento de ciertos objetos usando variables aleatorias y distribuciones de probabilidad. Estos modelos se usan ampliamente para representar datos y sus categorías.

Por ejemplo:

- En tareas de minería de datos como caracterización o clasificación, los modelos estadísticos pueden ser el resultado final.
- También la minería de datos puede apoyarse en modelos estadísticos, como cuando se usan para manejar ruido o valores faltantes en los datos.

La investigación en estadística desarrolla herramientas para predicción y pronóstico, basadas en modelos estadísticos. Además:

- Sirve para resumir y describir datos.
- Ayuda a extraer patrones y entender los mecanismos que los generan.
- Con la estadística inferencial, se modelan la aleatoriedad e incertidumbre para sacar conclusiones sobre un proceso o población.

Otro punto clave es que la estadística puede validar los resultados de la minería de datos.

Ejemplo:

- Si construimos un modelo de clasificación o predicción, este debe verificarse mediante pruebas de hipótesis estadísticas.
- Un resultado es estadísticamente significativo si es poco probable que ocurra por azar. Esto fortalece la confiabilidad del modelo.

Sin embargo, aplicar métodos estadísticos en minería de datos no es tarea sencilla:

- Muchos métodos estadísticos tienen una alta complejidad computacional.
- Escalarlos a grandes volúmenes de datos distribuidos requiere diseñar algoritmos optimizados.
- El reto es aún mayor en aplicaciones en tiempo real, como sugerencias de búsqueda online, donde se deben procesar flujos de datos continuos y veloces.

La investigación en minería de datos ha desarrollado soluciones escalables y eficientes para analizar grandes conjuntos de datos y flujos en tiempo real. Según el tipo de datos y la aplicación, se requieren métodos de análisis distintos. Gracias a esto, han surgido nuevas técnicas estadísticas escalables basadas en minería de datos.

5.2. Aprendizaje automático y mineria de mineria de datos

El aprendizaje automático (machine learning) estudia cómo las computadoras pueden aprender o mejorar su desempeño a partir de los datos.

Es una disciplina que crece muy rápido, con muchas metodologías y aplicaciones recientes, como:

- Máquinas de vectores de soporte (SVM),
- Modelos gráficos probabilísticos,
- Aprendizaje profundo (*deep learning*),

  y más.

En general, el machine learning aborda dos problemas clásicos:

Aprendizaje supervisado

Ejemplo clásico: clasificación.

- El aprendizaje se guía por ejemplos etiquetados en los datos de entrenamiento.

  Caso: para reconocer automáticamente códigos postales escritos a mano en cartas, el sistema aprende con un conjunto de imágenes de códigos escritos a mano junto con su traducción a números legibles por máquina. A partir de eso, construye un modelo de clasificación.

Aprendizaje no supervisado

Ejemplo clásico: clustering (agrupamiento).

- El proceso no es supervisado porque los datos no tienen etiquetas.

  Caso: un sistema recibe como entrada un conjunto de imágenes de dígitos escritos a mano. Si encuentra 10 grupos, es probable que correspondan a los dígitos del 0 al 9. Sin embargo, al no haber etiquetas, el modelo no puede decir con certeza qué número representa cada grupo.

5.3. Tecnología de base de datos y minería de datos

La investigación en sistemas de bases de datos se centra en la creación, mantenimiento y uso de bases de datos para organizaciones y usuarios finales. En este campo, se han desarrollado principios bien establecidos en:

- Modelos de datos,
- Lenguajes de consulta,
- Procesamiento y optimización de consultas,
- Métodos de almacenamiento de datos,
- Métodos de indexación.

La tecnología de bases de datos es muy reconocida por su escalabilidad, es decir, la capacidad de procesar de manera eficiente conjuntos de datos muy grandes y estructurados.

Por otro lado, muchas tareas de minería de datos requieren trabajar con grandes volúmenes de información o incluso con flujos de datos en tiempo real. Aquí, la minería de datos puede aprovechar la tecnología de bases de datos para lograr eficiencia y escalabilidad.

Además, la minería de datos no solo se beneficia de las bases de datos, sino que también extiende sus capacidades, permitiendo cumplir con las necesidades de análisis de datos más complejas de los usuarios.

En la actualidad, los sistemas de bases de datos han incorporado capacidades sistemáticas de análisis, gracias a:

- Almacenes de datos (Data Warehousing)
- Herramientas de minería de datos integradas

Un almacén de datos integra información de múltiples fuentes y distintos períodos de tiempo, y la organiza en un espacio multidimensional formando lo que se conoce como cubos de datos (*data cubes*).

Este modelo de cubos de datos facilita el procesamiento analítico en línea (OLAP) en bases de datos multidimensionales, y además promueve el desarrollo de la minería de datos multidimensional, que se estudiará más adelante.

5.4. Minería de datos y otras disciplinas

Además de su relación con la estadística, el aprendizaje automático y la tecnología de bases de datos, la minería de datos también se conecta estrechamente con muchas otras disciplinas.

La mayoría de los datos del mundo real son no estructurados, es decir, se presentan como texto en lenguaje natural, imágenes, audio o video.
Por eso, disciplinas como:

- Procesamiento de lenguaje natural (NLP),

- Visión por computadora,
- Reconocimiento de patrones,
- Procesamiento de señales de audio y video,
- Recuperación de información,

juegan un papel clave para poder manejar este tipo de datos.

Además, cada tipo especial de datos requiere integrar conocimiento del dominio correspondiente en el diseño de algoritmos de minería. Por ejemplo:

- Minería de datos biomédicos → necesita integrar biología, medicina e bioinformática.
- Minería de datos geoespaciales → requiere geografía y ciencias de datos espaciales.
- Minería de errores en programas de software → combina ingeniería de software con minería de datos.
- Minería en redes sociales o medios sociales → necesita apoyo de ciencias sociales y ciencias de redes.

En realidad, la minería de datos se extiende a casi todos los dominios de aplicación.

Desafíos principales

Uno de los grandes retos de la minería de datos es la eficiencia y escalabilidad, ya que a menudo se manejan cantidades enormes de datos bajo fuertes limitaciones de tiempo y recursos.

Por eso, la minería de datos está fuertemente conectada con el diseño de algoritmos eficientes, tales como:

- Algoritmos de baja complejidad,
- Algoritmos incrementales,
- Algoritmos para minería de flujos de datos (streaming).

También depende de:

- Computación de alto rendimiento,
- Computación paralela y distribuida,
- Computación en la nube y en clústeres.

Interacción humano-computadora (HCI)

La minería de datos también está relacionada con la interacción humano-computadora, ya que los usuarios necesitan:

- Indicar al sistema qué datos minar,
- Incorporar conocimientos previos en el proceso,
- Decidir cómo minar,
- Recibir los resultados de una forma fácil de entender (ej. visualizaciones, interpretaciones),
- Usar interfaces amigables e interactivas (ej. interfaces gráficas intuitivas).

Actualmente, existen muchos sistemas de minería de datos interactivos y aún más aplicaciones que ocultan funciones de minería de datos dentro de sus programas.

No se espera que toda la sociedad entienda o domine las técnicas de minería de datos, y además, las empresas no pueden exponer todos sus datos abiertamente. Por eso, muchas aplicaciones ya incluyen minería de datos integrada de forma casi invisible.

Ejemplos:

- Motores de búsqueda inteligentes,
- Tiendas online que usan el historial de búsqueda o compra de los usuarios para dar recomendaciones personalizadas.

6. Minería de datos y aplicaciones

Donde hay datos, hay aplicaciones de minería de datos.
La minería de datos es una disciplina muy orientada a la aplicación y ha tenido éxito en muchos sectores.

A continuación, algunos ejemplos destacados:

6.1. Business Intelligence (BI)

Las empresas necesitan entender mejor a sus clientes, mercados, proveedores, recursos y competidores.

BI ofrece visión histórica, actual y predictiva de las operaciones.

Herramientas:

- Reportes

- OLAP (procesamiento analítico en línea)
- Gestión del rendimiento empresarial
- Inteligencia competitiva
- Benchmarking
- Analítica predictiva

Rol de la minería de datos en BI

- ❖ Es el núcleo de la inteligencia de negocios.
- ❖ Sin minería de datos, sería muy difícil hacer análisis de mercado, comparar feedback de clientes o descubrir debilidades de competidores.

Técnicas usadas:

- ❖ Clasificación y predicción → analítica predictiva.
- ❖ Clustering → gestión de relaciones con clientes (segmentación).
- ❖ Resúmenes multidimensionales → entender mejor grupos de clientes y crear programas de fidelización personalizados.

6.2. Motores de búsqueda (Search Engines)

- ❖ Devuelven resultados de consultas como listas de páginas, imágenes o archivos.
- ❖ Diferencia con directorios web: los motores se basan en algoritmos, no solo en editores humanos.

Retos de minería de datos en buscadores:

I. Escalabilidad → manejo de enormes volúmenes de datos distribuidos en la nube.

II. Datos en línea → modelos deben entrenarse offline pero adaptarse rápidamente online (ejemplo: clasificar “apple” como fruta o marca).
III. Actualización incremental → flujos de datos crecientes y consultas nuevas cada momento.
IV. Consultas raras → la mayoría de queries son únicas o poco frecuentes, lo que dificulta recomendaciones contextuales.
V. Redes sociales y medios sociales

La minería de datos aquí permite analizar grandes cantidades de datos generados por interacciones sociales.

7 Minería de datos y sociedad

En 2023, la Unión Europea sancionó a Meta (Facebook/Instagram) con 1,200 millones de euros por transferir datos de usuarios europeos a Estados Unidos sin cumplir con las normas de protección de datos GDPR(Reglamento General de Protección de Datos).

- Este caso evidenció el mal uso de la minería de datos y el manejo indebido de información personal a gran escala.
- Mostró cómo el incumplimiento en la protección de datos puede tener consecuencias legales y económicas severas, además de generar desconfianza social.

8 Resumen

- **Necesidad y origen del Data Mining**

Surge como respuesta al gran crecimiento de datos. Es la evolución natural de la tecnología de la información y combina varias disciplinas.

- **Definición de Data Mining**

  Es descubrir patrones y conocimiento en grandes volúmenes de datos, mediante un proceso que incluye limpieza, integración, selección, transformación, descubrimiento, evaluación y presentación.

- **Patrones interesantes**

  Un patrón es valioso si es válido, novedoso, útil y fácil de entender. Estos patrones representan conocimiento y pueden medirse de forma objetiva o subjetiva.

- **Tipos de datos**

  Puede aplicarse a datos estructurados (bases de datos, transacciones) o no estructurados (texto, multimedia). También a datos almacenados o en flujo (stream).

- **Funcionalidades del Data Mining**

  Incluye caracterización, discriminación, patrones frecuentes, asociaciones, correlaciones, clasificación, regresión, deep learning, clustering y detección de anomalías.

- **Relación con otras disciplinas**

  Se conecta con estadística, machine learning, bases de datos, entre otras, aunque tiene un enfoque propio de investigación.

- **Aplicaciones**

Se aplica en inteligencia de negocios, búsquedas web, bioinformática, salud, finanzas, bibliotecas digitales y gobiernos digitales.

- **Impacto social**

  Tiene un fuerte impacto en la sociedad y es necesario estudiar cómo garantizar su efectividad y un uso responsable.

9 Ejercicios

1.1. ¿Qué es la minería de datos? En tu respuesta, aborda lo siguiente:

Definición: Es el proceso de descubrir patrones y conocimiento útil a partir de grandes volúmenes de datos.

a. ¿Es una simple transformación o aplicación de la tecnología desarrollada a partir de bases de datos, estadística, aprendizaje automático y reconocimiento de patrones?

No es solo una simple aplicación de bases de datos, estadística, aprendizaje automático o reconocimiento de patrones, sino la integración de todas estas áreas.

b. Alguien cree que la minería de datos es un resultado inevitable de la evolución de la tecnología de la información. Si fueras un investigador en bases de datos, muestra que la minería de datos resulta de una evolución natural de la tecnología de bases de datos. ¿Y si fueras un investigador en aprendizaje automático o un estadístico?

- Como investigador en bases de datos, se entiende como evolución natural: de simples consultas y almacenamiento a la extracción de conocimiento.

- Como experto en aprendizaje automático, la minería de datos es una aplicación masiva de modelos predictivos y algoritmos.
- Como estadístico, es una extensión moderna del análisis de datos tradicional aplicada a grandes volúmenes y distintos formatos.

c. Describe los pasos involucrados en la minería de datos cuando se la considera como un proceso de descubrimiento de conocimiento.

Etapas: limpieza, integración, selección, transformación, descubrimiento de patrones/modelos, evaluación y presentación del conocimiento.

1.2. Define cada una de las siguientes funcionalidades de la minería de datos: análisis de asociación y correlación, clasificación, regresión, agrupamiento (clustering), aprendizaje profundo (deep learning) y análisis de valores atípicos (outlier analysis). Da ejemplos de cada funcionalidad de minería de datos, usando una base de datos de la vida real con la que estés familiarizado.

Funcionalidades de la minería de datos

- Asociación y correlación: Encuentra relaciones entre ítems. Ejemplo: en supermercados, identificar que "pan + mantequilla → leche" (reglas de mercado).
- Clasificación: Asigna objetos a clases predefinidas. Ejemplo: aprobar o rechazar créditos bancarios.
- Regresión: Predice valores numéricos. Ejemplo: estimar el precio de una vivienda.
- Clustering (agrupamiento): Agrupa objetos similares sin categorías predefinidas. Ejemplo: segmentación de clientes.
- Deep Learning: Detecta patrones complejos en imágenes, texto o voz. Ejemplo: reconocimiento facial en seguridad.

- Análisis de outliers: Detecta valores atípicos. Ejemplo: identificar transacciones fraudulentas en tarjetas de crédito.

1.3. Presenta un ejemplo donde la minería de datos sea crucial para el éxito de un negocio. ¿Qué funcionalidades de minería de datos necesita este negocio (por ejemplo, piensa en los tipos de patrones que podrían extraerse)? ¿Pueden generarse tales patrones de manera alternativa mediante procesamiento de consultas de datos o un simple análisis estadístico?

En Rappi o Uber Eats, la minería de datos es crucial para recomendar restaurantes, optimizar rutas y predecir demanda en horas pico.

- Funcionalidades necesarias: patrones frecuentes, clustering (segmentación de clientes), clasificación (predicción de pedidos), regresión (tiempo de entrega estimado).
- Comparación: Procesamiento de consultas o estadística simple no puede encontrar patrones ocultos o predecir con precisión, por lo que la minería es esencial.

1.4. Explica la diferencia y similitud entre:

- **análisis de correlación y clasificación,**
- **clasificación y agrupamiento,**
- **clasificación y regresión.**

Diferencias y similitudes

- Correlación vs Clasificación:
    - Correlación: mide relaciones entre variables.
    - Clasificación: asigna categorías según atributos.

    - Similitud: ambas buscan dependencias entre datos.
- Clasificación vs Clustering:
    - Clasificación: requiere clases predefinidas.
    - Clustering: agrupa sin clases previas.
    - Similitud: ambas organizan datos según características.
- Clasificación vs Regresión:
    - Clasificación: predice clases categóricas (ej. aprobar/no aprobar).
    - Regresión: predice valores continuos (ej. precio, tiempo).

1.5. Basado en tus observaciones, describe otro posible tipo de conocimiento que deba descubrirse mediante métodos de minería de datos, pero que no ha sido listado en este capítulo.

¿Requiere una metodología de minería bastante diferente a las descritas en este capítulo?

Otro conocimiento posible a descubrir

- Detección de emociones en texto/redes sociales.
    - Ejemplo: analizar sentimientos de clientes para medir reputación de marca.
    - Requiere técnicas de minería de texto y PLN (Procesamiento de Lenguaje Natural).

1.6. Los valores atípicos a menudo se descartan como ruido. Sin embargo, la basura de una persona puede ser el tesoro de otra.

Por ejemplo, las excepciones en transacciones con tarjetas de crédito pueden ayudarnos a detectar el uso fraudulento de estas. Usando la detección de fraudes como ejemplo, propone dos métodos que pueden usarse para detectar valores atípicos y discute cuál es más confiable.

Outliers (valores atípicos)

- Métodos de detección:
    - Basado en reglas: umbrales definidos (ej. transacciones de monto muy alto).
    - Métodos estadísticos o de clustering: identificar puntos que no encajan en grupos normales.
- Más confiable: los métodos basados en clustering y modelos estadísticos, porque se adaptan al contexto en vez de depender de reglas fijas.

1.7. ¿Cuáles son los principales desafíos de minar una gran cantidad de datos (por ejemplo, miles de millones de tuplas) en comparación con minar una pequeña cantidad de datos (por ejemplo, un conjunto de unos pocos cientos de tuplas)?

Desafíos de grandes vs pequeños volúmenes

- **Grandes volúmenes:** problemas de escalabilidad, eficiencia, almacenamiento, tiempo de procesamiento y necesidad de algoritmos distribuidos.
- **Pequeños volúmenes:** más fácil de manejar, pero menor poder predictivo y patrones menos representativos.

1.8. Describe los principales desafíos de investigación de la minería de datos en una aplicación específica.

Ejemplo: Salud digital.

- Desafíos: privacidad de datos médicos, integración de distintas fuentes (hospitales, laboratorios, wearables), y precisión en predicciones críticas (diagnóstico).
- Se requiere minería que preserve la privacidad y que sea altamente confiable para la toma de decisiones médicas.