¿Qué es la minería de datos?

La minería de datos es el uso del machine learning y el análisis estadístico para descubrir patrones y otra información valiosa a partir de grandes conjuntos de datos.

Dada la evolución del machine learning (ML), el [almacenamiento de datos](https://www.ibm.com/es-es/topics/data-warehouse) y el crecimiento del [big data](https://www.ibm.com/es-es/topics/big-data-analytics" \t "_self), la adopción de la minería de datos, también conocida como descubrimiento de conocimientos en bases de datos (KDD), se ha acelerado rápidamente en las últimas décadas. Sin embargo, aunque esta tecnología evoluciona continuamente para gestionar los datos a gran escala, los líderes podrían enfrentarse a desafíos con la escalabilidad y la [automatización](https://www.ibm.com/es-es/topics/automation).

Las técnicas de minería de datos que sustentan los análisis de datos se pueden implementar para dos propósitos principales. Pueden describir el conjunto de datos objetivo o pueden predecir resultados mediante el uso de algoritmos de machine learning.

Estos métodos se utilizan para organizar y filtrar los datos, sacando a la luz la información más útil, desde el fraude hasta los comportamientos de los usuarios, los cuellos de botella e incluso las violaciones de seguridad. El uso de algoritmos de ML e [inteligencia artificial (IA)](https://www.ibm.com/es-es/topics/artificial-intelligence) permite la automatización del análisis, lo que puede acelerar enormemente el proceso.

Cuando se combina con herramientas de [análisis y](https://www.ibm.com/es-es/blog/data-science-vs-data-analytics-unpacking-the-differences/) [visualización](https://www.ibm.com/es-es/topics/data-visualization) de datos, como [Apache Spark](https://www.ibm.com/es-es/topics/apache-spark), el software de minería de datos es cada vez más sencillo y la extracción de información relevante se puede obtener más rápido que nunca. Los avances en IA continúan acelerando la adopción en todas las industrias.

[Libro electrónico](https://www.ibm.com/account/reg/signup?formid=urx-52356" \t "_blank)

[IA generativa y ML para la empresa](https://www.ibm.com/account/reg/signup?formid=urx-52356" \t "_blank)

[Aprenda a incorporar con confianza la IA generativa y el machine learning a su negocio.](https://www.ibm.com/account/reg/signup?formid=urx-52356" \t "_blank)

[Regístrese para descargarlo](https://www.ibm.com/account/reg/signup?formid=urx-52356)

Beneficios y desafíos

**Beneficios**

**Descubra ideas y tendencias ocultas**: la minería de datos toma los datos en bruto y encuentra el orden en el caos: ver el bosque por los árboles. Esto puede dar lugar a una planificación mejor informada en todos los sectores y funciones corporativas, incluyendo publicidad, finanzas, [gobierno](https://www.ibm.com/es-es/case-studies/state-gov-process-mining), sanidad, recursos humanos, fabricación, marketing, investigación, ventas y [gestión de la cadena de suministro (SCM)](https://www.ibm.com/es-es/topics/supply-chain-logistics).

**Ahorre presupuesto**: al analizar los datos de rendimiento de múltiples fuentes, se pueden identificar los cuellos de botella en los procesos empresariales para acelerar su resolución y aumentar la eficacia.

**Resuelva múltiples desafíos**: la minería de datos es una herramienta versátil. Los datos de casi cualquier fuente y cualquier aspecto de una organización se pueden analizar para descubrir patrones y mejores formas de hacer negocios. Casi todos los departamentos de una organización que recopilan y [analizan datos](https://www.ibm.com/es-es/topics/exploratory-data-analysis) pueden beneficiarse de la minería de datos.

**Desafíos**

**Complejidad y riesgo**: la información útil requiere datos válidos, además de expertos con experiencia en codificación. El conocimiento de los lenguajes de minería de datos, incluidos Python, R y SQL, es útil. Un enfoque insuficientemente cauteloso de la minería de datos podría provocar resultados engañosos o peligrosos. Algunos datos de consumidores utilizados en la minería de datos pueden ser [información de identificación personal (PII)](https://www.ibm.com/es-es/topics/pii) que se deben gestionar con cuidado para evitar problemas legales o de relaciones públicas.

**Coste**: para obtener los mejores resultados, a menudo se necesita una colección amplia y profunda de conjuntos de datos. Si una organización va a recopilar nueva información, crear una canalización de datos puede representar un gasto nuevo. Si los datos deben comprarse de una fuente externa, eso también impone un coste.

**Incertidumbre**: en primer lugar, un gran esfuerzo de minería de datos puede estar bien gestionado, pero producir resultados poco claros, sin ningún beneficio importante. O los datos inexactos pueden dar lugar a percepciones incorrectas, ya sea porque se seleccionaron datos incorrectos o porque el preprocesamiento se realizó de forma incorrecta. Otros riesgos incluyen errores de modelado o datos obsoletos de un mercado que cambia rápidamente.

Otro problema potencial es que los resultados pueden *parecer* válidos pero en realidad son aleatorios y no son de fiar. Es importante recordar que la "correlación no es causal". Un famoso ejemplo de "dragado de datos", ver una correlación aparente y exagerar su importancia, fue presentado recientemente por el bloguero Tyler Vigen: "El precio de las acciones de Amazon.com coincide estrechamente con el número de niños llamados 'Stevie' entre 2002 y 2022".1 Pero, por supuesto, el nombre de Stevies no influyó en el precio de las acciones ni viceversa. Las aplicaciones de minería de datos encuentran los patrones, pero el juicio humano sigue siendo importante.

Minería de datos versus minería de texto versus minería de procesos

**La minería de datos** es el proceso general de identificar patrones y extraer información útil de grandes conjuntos de datos. Se puede utilizar para evaluar datos estructurados y no estructurados a fin de identificar información nueva y se utiliza normalmente para analizar el comportamiento de los consumidores para los equipos de marketing y ventas. Por ejemplo, los métodos de extracción de datos se pueden utilizar para observar y predecir comportamientos, incluidoas la pérdida de clientes, la [detección de fraudes](https://www.ibm.com/es-es/topics/fraud-detection), el análisis de canasta de mercado y más.

**La minería de textos**, también conocida como minería de datos de texto, es un subcampo de la minería de datos, destinado a transformar texto no estructurado en un formato estructurado para identificar patrones significativos y generar nuevos conocimientos. Los datos no estructurados pueden incluir texto de fuentes, como publicaciones en redes sociales, reseñas de productos, artículos, correo electrónico o formatos de medios enriquecidos, como archivos de vídeo y audio. Gran parte de los datos disponibles públicamente en todo el mundo no están estructurados, lo que hace que [la minería de textos](https://www.ibm.com/es-es/topics/text-mining) sea una práctica valiosa.

**La minería de procesos** se encuentra en la intersección de la [gestión de procesos empresariales (BPM)](https://www.ibm.com/es-es/topics/business-process-management) y la minería de datos. [La minería de procesos](https://www.ibm.com/es-es/topics/process-mining) proporciona una forma de aplicar algoritmos a los datos del registro de eventos para identificar tendencias, patrones y detalles de cómo se desarrollan los procesos. La minería de procesos aplica [la ciencia de datos](https://www.ibm.com/es-es/topics/data-science) para descubrir cuellos de botella y luego validar y mejorar [los flujos de trabajo](https://www.ibm.com/es-es/topics/workflow).

Por lo general, el BPM recopila datos de manera más informal a través de talleres y entrevistas y, a continuación, utiliza programas informáticos para documentar ese flujo de trabajo en forma de [mapa de procesos](https://www.ibm.com/es-es/topics/process-mapping). Dado que los datos que sirven de base a estos mapas de procesos suelen ser cualitativos, la minería de procesos aporta un enfoque más cuantitativo a un problema de procesos, detallando el proceso real a través de datos de eventos.

Los sistemas de información, como las [herramientas de planificación de recursos empresariales (ERP)](https://www.ibm.com/es-es/topics/enterprise-resource-planning) o [de gestión de relaciones con los clientes (CRM),](https://www.ibm.com/es-es/topics/crm) proporcionan una pista de auditoría de los procesos con sus respectivos datos de registro. La minería de procesos utiliza estos datos de los sistemas de TI para ensamblar un modelo de proceso o un gráfico de procesos. A partir de ahí, las organizaciones pueden examinar el proceso de principio a fin con los detalles y las variaciones descritas.

Cómo funciona la minería de datos

El proceso de extracción de datos implica varios pasos, desde la recopilación de datos hasta la [visualización](https://www.ibm.com/es-es/blog/data-virtualization-unifies-data-for-seamless-ai-and-analytics/), para extraer información valiosa de grandes conjuntos de datos. Las técnicas de extracción de datos se pueden utilizar para generar descripciones y predicciones sobre un conjunto de datos objetivo.

Los científicos de datos o especialistas en [inteligencia empresarial (BI)](https://www.ibm.com/es-es/topics/business-intelligence) describen los datos a través de sus observaciones de patrones, asociaciones y correlaciones. También clasifican y agrupan datos mediante métodos de clasificación y regresión, e identifican valores atípicos para casos de uso, como la detección de spam.

La minería de datos suele incluir cinco pasos principales: establecimiento de objetivos, selección de datos, preparación de datos, creación de modelos de datos y minería de patrones y evaluación de resultados.

**1. Establecer los objetivos empresariales:**esta puede ser la parte más difícil del proceso de minería de datos, y muchas organizaciones dedican muy poco tiempo a este importante paso. Incluso antes de identificar, extraer o limpiar los datos, los [científicos de datos](https://www.ibm.com/es-es/topics/data-science) y las partes interesadas del negocio pueden trabajar juntas para definir el problema empresarial preciso, lo que ayuda a informar las preguntas y parámetros de los datos de un proyecto. Es posible que los analistas también deban investigar más para comprender completamente el contexto empresarial.

**2. Selección de datos:**cuando se define el alcance del problema, es más fácil para los científicos de datos identificar qué conjunto de datos ayudará a responder las preguntas pertinentes para el negocio. Ellos y el equipo de TI también pueden determinar dónde se deben almacenar y proteger los datos.

**3. Preparación de datos:** los datos relevantes se recopilan y [limpian](https://www.ibm.com/es-es/topics/data-profiling) para eliminar cualquier ruido, como duplicados, valores faltantes y valores atípicos. En función del conjunto de datos, se puede tomar un paso adicional de [gestión de datos](https://www.ibm.com/es-es/topics/data-management) para reducir el número de dimensiones, ya que demasiadas entidades pueden ralentizar cualquier cálculo posterior.

Los científicos de datos buscan conservar los predictores más importantes para ayudar a garantizar una [precisión óptima](https://www.ibm.com/es-es/blog/3-new-steps-in-the-data-mining-process-to-ensure-trustworthy-ai/) dentro de cualquier modelo. Una ciencia de datos responsable implica pensar en el modelo más allá del código y el rendimiento, y se ve muy afectado por los datos que se utilizan y por su fiabilidad.

**4. Creación de modelos y minería de patrones:** según el tipo de análisis, los científicos de datos pueden investigar cualquier tendencia o relación de datos interesante, como patrones secuenciales, reglas de asociación o correlaciones. Aunque los patrones de alta frecuencia tienen aplicaciones más amplias, a veces las desviaciones en los datos pueden ser más interesantes y resaltar áreas de posible fraude. [Los modelos predictivos](https://www.ibm.com/es-es/topics/predictive-analytics) pueden ayudar a evaluar tendencias o resultados futuros. En los sistemas más sofisticados, los modelos predictivos pueden realizar predicciones en tiempo real para obtener respuestas rápidas a los mercados cambiantes.

Los algoritmos [de deep learning](https://www.ibm.com/es-es/topics/deep-learning) también podrían usarse para clasificar o agrupar un conjunto de datos en función de los datos disponibles. Si los datos de entrada están etiquetados (como en el [aprendizaje supervisado](https://www.ibm.com/es-es/topics/supervised-learning)), se puede utilizar un modelo de clasificación para categorizar datos o, alternativamente, se puede aplicar una regresión para predecir la probabilidad de una asignación determinada. Si el conjunto de datos no está etiquetado (es decir, [aprendizaje no supervisado](https://www.ibm.com/es-es/topics/unsupervised-learning)), los puntos de datos individuales en el conjunto de entrenamiento se comparan para descubrir similitudes subyacentes y se agrupan en función de esas características.

**5. Evaluación de los resultados e implementación del conocimiento:** cuando los datos se agregan, se pueden preparar para su presentación, a menudo mediante técnicas de [visualización de datos](https://www.ibm.com/es-es/topics/data-visualization), de modo que los resultados se puedan evaluar e interpretar. Lo ideal es que los resultados finales sean válidos, novedosos, útiles y comprensibles. Cuando se cumplen estos criterios, los responsables de la toma de decisiones pueden utilizar este conocimiento para implementar nuevas estrategias, logrando los objetivos previstos.

Técnicas de minería de datos

Estos son algunos de los tipos más populares de minería de datos:

**Reglas de la asociación:**una regla de asociación es un método basado en reglas si/entonces para encontrar relaciones entre variables en un conjunto de datos. La fuerza de las relaciones se mide por el apoyo y la confianza. El nivel de confianza se basa en la frecuencia con la que las sentencias si o entonces son verdaderas. La medida de soporte es la frecuencia con la que se muestran los elementos relacionados en los datos.

Estos métodos se utilizan con frecuencia para el análisis de la cesta de la compra, lo que permite a las empresas comprender mejor las relaciones entre diferentes productos, como los que se compran juntos con frecuencia. Comprender los hábitos de los clientes permite a las empresas desarrollar mejores estrategias de venta cruzada y motores de recomendación.  
 **Clasificación**: se predefinen clases de objetos, según las necesidades de la organización, con definiciones de las características que los objetos tienen en común. Esto permite agrupar los datos subyacentes para facilitar su análisis.

Por ejemplo, una empresa de productos de consumo podría examinar su estrategia de cupones revisando los reembolsos de cupones anteriores junto con los datos de ventas, las estadísticas de inventario y cualquier dato de consumo disponible para encontrar la mejor estrategia de campaña futura.  
 **Agrupación**: estrechamente relacionada con la clasificación, la agrupación informa de similitudes, pero también proporciona más agrupaciones basadas en diferencias. Las clasificaciones preestablecidas para un fabricante de jabón pueden incluir detergente, lejía, suavizante de ropa, limpiador de pisos y cera para pisos; mientras que la agrupación en clústeres puede crear grupos que incluyen productos de lavandería y cuidado del suelo.  
 **Árbol de decisión:**esta técnica de minería de datos utiliza el análisis de clasificación o regresión para clasificar o predecir posibles resultados en función de un conjunto de decisiones. Como sugiere el nombre del [árbol de decisión](https://www.ibm.com/es-es/topics/decision-trees), utiliza una visualización similar a un árbol para representar los posibles resultados de estas decisiones.

**Vecino K más cercano (KNN):** también conocido como algoritmo KNN, [vecino K más cercano](https://www.ibm.com/es-es/topics/knn) es un algoritmo no paramétrico que clasifica los puntos de datos según su proximidad y asociación con otros datos disponibles. Este algoritmo asume que se encuentran puntos de datos similares cerca unos de otros. Como resultado, trata de calcular la distancia entre los puntos de datos, generalmente a través de la distancia euclidiana, y luego asigna una categoría basada en la categoría o media más frecuente.  
 **Redes neuronales:**utilizadas principalmente para algoritmos de deep learning, [redes neuronales](https://www.ibm.com/es-es/topics/neural-networks) procesan datos por medio de los algoritmos de conectividad. Cada nodo se compone de entradas, ponderaciones, un sesgo y una salida.

Si ese valor de salida supera un umbral determinado, se "dispara" o activa el nodo, pasando los datos a la siguiente capa de la red. Las redes neuronales aprenden esta función de mapeo a través del aprendizaje supervisado, realizando ajustes basados en la función de pérdida a través del proceso de descenso gradiente. Cuando la función de costes es igual o cercana a cero, una organización puede confiar en la precisión del modelo para dar con la respuesta correcta.

**Análisis predictivo:** Alcombinar la minería de datos con técnicas de modelado estadístico y machine learning, se pueden analizar datos históricos utilizando el análisis predictivo para crear modelos gráficos o matemáticos destinados a identificar patrones, prever acontecimientos y resultados futuros e identificar riesgos y oportunidades.  
 **Análisis de regresión**:esta técnica descubre relaciones en los datos mediante la predicción de resultados basados en variables predeterminadas. Esto puede incluir [árboles de decisión](https://www.ibm.com/es-es/topics/decision-trees) y [regresión lineal](https://www.ibm.com/es-es/topics/linear-regression) multivariada. Los resultados pueden priorizarse según la cercanía de la relación para ayudar a determinar qué datos son más o menos significativos. Un ejemplo sería que un fabricante de refrescos estimara el inventario necesario de bebidas antes de la llegada del caluroso verano previsto.

Casos de uso de minería de datos

Las técnicas de extracción de datos son ampliamente adoptadas por los equipos de [inteligencia empresarial](https://www.ibm.com/es-es/topics/business-intelligence) y análisis de datos, lo que les ayuda a extraer conocimientos para su organización y su sector. Algunos casos de uso de minería de datos incluyen:

**Detección de anomalías**aunque los patrones de datos frecuentes pueden proporcionar a los equipos información valiosa, observar [anomalías](https://www.ibm.com/es-es/topics/anomaly-detection) de datos también es beneficioso,ya que ayuda a las organizaciones con [detección de fraudes](https://www.ibm.com/es-es/topics/fraud-detection), intrusiones en la red y defectos de productos. Aunque se trata de un caso de uso bien conocido en la banca y otras instituciones financieras, las empresas basadas en SaaS también han empezado a adoptar estas prácticas para eliminar las cuentas de usuario falsas de sus conjuntos de datos. La detección de anomalías también puede ser una oportunidad para encontrar estrategias nuevas y novedosas o mercados objetivo que se hayan pasado por alto en el pasado.

**Evaluación del riesgo**  
Las organizaciones pueden localizar y determinar la magnitud del riesgo con mayor precisión con la minería de datos. Se pueden descubrir patrones y anomalías en los campos de la [ciberseguridad](https://www.ibm.com/es-es/topics/cybersecurity), las finanzas y el ámbito legal para identificar descuidos o amenazas.

**Atención a los mercados objetivo**  
Al buscar en varias bases de datos para encontrar relaciones cercanas, la minería de datos puede conectar con precisión los comportamientos y los orígenes de los clientes con las ventas de artículos específicos. Esto puede permitir campañas más específicas para ayudar a impulsar las ventas.

**Mejora del servicio de atención al cliente**  
Los problemas de los clientes se pueden descubrir y solucionar antes si se puede revisar la suma total de las acciones del cliente (in situ, en línea, a través de aplicaciones móviles o por teléfono) mediante la minería de datos. Los agentes de servicio al cliente pueden tener acceso a información más completa y detallada sobre los clientes a los que atienden.  
  
**Aumento del tiempo de actividad del equipo**  
Los datos operativos se pueden extraer de equipos industriales que pueden ayudar a predecir el rendimiento y el tiempo de inactividad futuros y permiten planificar el mantenimiento de protección.  
  
**Optimización operativa**  
La minería de procesos utiliza técnicas de minería de datos para reducir los costes en las funciones operativas y permitir a las organizaciones funcionar de forma más eficiente. Esta práctica puede ayudar a identificar los costosos cuellos de botella y a mejorar la toma de decisiones para los líderes empresariales.

Casos de uso del sector

**Servicio de atención al cliente**  
La extracción de datos puede crear una fuente de datos más rica para el servicio de atención al cliente, ya que ayuda a determinar qué factores agradan más a los clientes y qué factores provocan fricción o insatisfacción.

**Educación**  
Los centros educativos han empezado a recopilar datos para entender su población estudiantil y qué entornos conducen al éxito. Dado que los cursos a menudo utilizan plataformas en línea, pueden usar varias dimensiones y métricas para observar y evaluar el rendimiento, como las pulsaciones de teclas, los perfiles de los estudiantes, las clases a las que asistieron y el tiempo empleado.

**Finanzas**  
Al investigar el riesgo, las instituciones financieras y los bancos suelen querer lanzar una amplia red para capturar cualquier factor que pueda afectar negativamente al flujo de caja y a la recuperación. Las herramientas de minería de datos pueden ser útiles para encontrar y sopesar una combinación de factores que indican un riesgo bueno o malo.

**Sanidad**  
La minería de datos es una herramienta útil para el diagnóstico de afecciones médicas, incluida la lectura de escaneos e imágenes, y luego ayuda a sugerir tratamientos beneficiosos.  
  
**Recursos humanos**Las organizaciones pueden obtener nuevos conocimientos sobre el rendimiento y la satisfacción de los empleados analizando múltiples factores y encontrando patrones. Los datos pueden incluir la fecha de inicio, el mandato, los ascensos, el salario, la formación, el desempeño de los compañeros, el desempeño del trabajo, el uso de las Capacidades y los viajes.  
  
**Fabricación**  
Desde las materias primas hasta la entrega final, todos los aspectos del proceso de fabricación se pueden analizar para mejorar el rendimiento. ¿Cuál es el coste de los materiales y hay opciones? ¿Cuál es la eficiencia de la producción? ¿Dónde están los cuellos de botella? ¿Cuáles son los problemas de calidad y dónde surgen, tanto internamente como con los clientes?

**Venta minorista**Al analizar los datos y las acciones de los clientes, los minoristas pueden identificar las campañas, los precios, las promociones, las ofertas de productos especiales y las ventas cruzadas y adicionales más productivas.

**Ventas y marketing**  
Las empresas recopilan cantidades masivas de datos sobre sus clientes y prospectos. Al observar la demografía de los consumidores, las respuestas de los medios de comunicación y el comportamiento de los clientes, las empresas pueden utilizar los datos para optimizar sus campañas de marketing, mejorar la segmentación y la orientación y los programas de fidelización de clientes, todo lo cual ayuda a obtener un mayor retorno de la inversión (ROI) en los esfuerzos de marketing. [Los análisis predictivos](https://www.ibm.com/es-es/topics/predictive-analytics) también pueden ayudar a los equipos a establecer expectativas con sus partes interesadas, proporcionando estimaciones de rendimiento para cualquier aumento o disminución de la inversión en marketing.

**Redes sociales**  
El análisis de los datos de los usuarios puede ayudar a descubrir nuevas oportunidades editoriales o nuevas fuentes de ingresos publicitarios para audiencias objetivo específicas.

**Gestión de la cadena de suministro (SCM)**  
Mediante la minería de datos, los gestores de productos pueden predecir mejor la demanda, preparar la producción, ajustar proveedores o adaptar los esfuerzos de marketing. Los administradores de la cadena de suministro pueden planificar mejor los envíos y el almacenamiento.