

**UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA**

**DESARROLLO DE SOFTWARE**



**EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS**

**REPORTE DE INVESTIGACIÓN DE LOS TIPOS DE APLICACIONES,  
PROCESAMIENTO Y HERRAMIENTAS PARA INTELIGENCIA ARTIFICIAL,  
MACHINE LEARNING, DATA MINING Y BIG DATA**

**PRESENTA:**

**MILDRED VILLASEÑOR RUIZ**

**DOCENTE:**

**ING. LUIS ENRIQUE MASCOTE CANO**

Chihuahua, Chih., 25 de septiembre de 2025

## Contenidos

|                                    |    |
|------------------------------------|----|
| Introducción .....                 | 3  |
| Inteligencia Artificial (IA) ..... | 3  |
| Definición .....                   | 3  |
| Aplicación .....                   | 3  |
| Procesamiento .....                | 3  |
| Batch VS Streaming .....           | 4  |
| Herramientas y tecnologías .....   | 4  |
| Machine Learning (ML) .....        | 5  |
| Definición .....                   | 5  |
| Aplicación .....                   | 5  |
| Procesamiento .....                | 5  |
| Batch VS Streaming .....           | 6  |
| Herramientas y tecnologías .....   | 6  |
| Data Mining (DM) .....             | 7  |
| Definición .....                   | 7  |
| Aplicación .....                   | 7  |
| Procesamiento .....                | 7  |
| Batch VS Streaming .....           | 8  |
| Herramientas y tecnologías .....   | 8  |
| Big Data .....                     | 9  |
| Definición .....                   | 9  |
| Aplicación .....                   | 9  |
| Procesamiento .....                | 9  |
| Batch VS Streaming .....           | 10 |
| Herramientas y tecnologías .....   | 10 |
| Conclusión .....                   | 11 |
| Referencias Bibliográficas .....   | 12 |

## Introducción

La Inteligencia Artificial, el Machine Learning, la Minería de Datos y el Big Data son tecnologías que cada vez escuchamos más porque se aplican en muchos aspectos de la vida diaria. Aunque suenen complicadas, en realidad su objetivo es usar datos para resolver problemas, hacer procesos más fáciles y ayudar en la toma de decisiones. Estas herramientas ya se usan en ejemplos comunes como las recomendaciones de películas, los chatbots o la detección de fraudes, y por eso es importante conocerlas y entender cómo influyen en diferentes áreas.

## Inteligencia Artificial (IA)

### Definición

La Inteligencia Artificial (IA) es la simulación de la inteligencia humana en máquinas para realizar tareas que normalmente requieren habilidades cognitivas humanas, como el aprendizaje, la resolución de problemas y la toma de decisiones.

### Aplicación

- Visión por Computador: Reconocimiento y análisis de imágenes o vídeos.

**Ejemplo:** Tesla emplea IA para que sus autos autónomos detecten peatones y señales de tránsito.

- Procesamiento de Lenguaje Natural (NLP): Capacidad de las máquinas para comprender y generar lenguaje humano.

**Ejemplo:** Los chatbots bancarios de BBVA responden automáticamente consultas de clientes.

### Procesamiento

1. Recolección y preparación de datos: Se recopilan cantidades masivas de datos relevantes para una tarea específica.
2. Entrenamiento de algoritmos: Los algoritmos de aprendizaje automático (machine learning) son entrenados con estos datos para que aprendan a reconocer patrones y a ajustar sus parámetros para minimizar errores.

3. Desarrollo de modelos: Se crean modelos de IA que imitan la forma en que los humanos aprenden, razonan y toman decisiones.
4. Análisis y predicción: Los modelos procesan la información para generar respuestas, hacer predicciones o automatizar tareas específicas.

### Batch VS Streaming

- Batch: El procesamiento por lotes (batch) en IA se refiere a la ejecución de tareas de forma masiva y no en tiempo real. Se usa para entrenar modelos con grandes volúmenes de datos, ya que no se necesita una respuesta inmediata.  
Ejemplo: El entrenamiento de un modelo de red neuronal para el reconocimiento de imágenes, donde se procesan miles o millones de fotos de una sola vez.
- Streaming: El procesamiento en tiempo real (streaming) implica el análisis de datos a medida que se generan, sin necesidad de almacenarlos primero. Es ideal para la inferencia, donde se requiere una respuesta instantánea.  
Ejemplo: Un sistema de detección de fraudes que analiza las transacciones de tarjeta de crédito en tiempo real para bloquear las sospechosas.

### Herramientas y tecnologías

1. Aprendizaje Automático (Machine Learning): Permite a los sistemas aprender de los datos, identificar patrones y hacer predicciones sin ser programados explícitamente para cada tarea.
2. Procesamiento del Lenguaje Natural (PLN): Facilita la comprensión e interacción entre las computadoras y el lenguaje humano, permitiendo la traducción, el resumen y la generación de texto.
3. Visión por Computadora: Capacita a las máquinas para "ver" e interpretar imágenes y videos, lo que es útil para el reconocimiento facial o la automatización de procesos visuales.
4. Redes Neuronales Profundas: Una subcategoría del machine learning que utiliza arquitecturas complejas para aprender patrones intrincados en grandes cantidades de datos, similar al funcionamiento del cerebro humano.

# Machine Learning (ML)

## Definición

Subcampo de la IA que diseña algoritmos que “aprenden” a partir de datos para realizar predicciones o tomar decisiones sin ser programados explícitamente para cada caso.

## Aplicación

- **Sistemas de Recomendación:** Algoritmos que predicen contenidos o productos que pueden interesar a un usuario.

**Ejemplo:** Netflix sugiere películas según el historial de cada suscriptor.

- **Detección de Fraude:** Identificación de transacciones sospechosas mediante patrones de comportamiento.

**Ejemplo:** MasterCard aplica modelos de ML para detectar fraudes en compras en tiempo real.

## Procesamiento

1. **Recopilación de Datos:** Obtener los datos brutos de diversas fuentes necesarios para el entrenamiento del modelo.
2. **Limpieza y Preparación:**
3. **Limpieza:** Manejar valores faltantes, datos inconsistentes y errores para asegurar la calidad de la información.
4. **Transformación:** Convertir los datos a un formato adecuado para el algoritmo, incluyendo la estandarización y normalización.
5. **Selección de Características:** Elegir las variables más relevantes para mejorar el rendimiento del modelo.
6. **Organización de Datos:**
7. **Almacenamiento:** Utilizar bases de datos (relacionales o NoSQL) o sistemas en la nube para gestionar grandes volúmenes de datos.
8. **Estructuración:** Organizar los datos en un formato comprensible para el modelo, como tablas o estructuras semi-estructuradas.

9. Elección del Modelo: Seleccionar el algoritmo de Machine Learning más adecuado (supervisado, no supervisado o por refuerzo) según el tipo de problema y los datos disponibles.
10. Entrenamiento del Modelo: Alimentar los datos procesados al algoritmo para que aprenda a identificar patrones y relaciones.
11. Evaluación del Modelo: Medir el rendimiento del modelo para asegurar su precisión y fiabilidad, a menudo mediante validación cruzada para evaluar su robustez.
12. Ajuste y Predicción: Ajustar los parámetros del modelo y realizar predicciones o tomar decisiones con mínima intervención humana, mejorando continuamente con nuevos datos.

### Batch VS Streaming

- Batch: Se utiliza en ML para tareas como el re-entrenamiento periódico de un modelo con nuevos datos. Por ejemplo, actualizar un modelo de predicción de ventas cada mes con los datos del mes anterior.
- Streaming: Se emplea en inferencia en tiempo real. Por ejemplo, un modelo de ML que analiza datos de sensores de un coche para predecir si un conductor está a punto de dormirse, requiriendo una respuesta inmediata.

### Herramientas y tecnologías

1. Python: Es el lenguaje más utilizado en Machine Learning, con un amplio ecosistema de librerías para análisis de datos, visualización y desarrollo de modelos.
2. TensorFlow: Desarrollada por Google, esta plataforma de código abierto es muy flexible y robusta, ideal para aplicaciones a gran escala.
3. PyTorch: Creada por Facebook, PyTorch destaca por su interfaz intuitiva y su rápida integración con Python, lo que facilita la experimentación y el prototipado rápido.
4. Scikit-learn: Una biblioteca de Python para el Machine Learning que incluye herramientas para clasificación, regresión, selección de modelos y clustering.
5. Keras: Una API para redes neuronales que puede ejecutarse sobre TensorFlow, facilitando la construcción de modelos complejos.
6. Pandas: Una librería fundamental para la manipulación y el análisis de datos en Python.

7. JAX: Una biblioteca de computación numérica de Google que facilita el trabajo con gradientes y la aceleración por hardware para la experimentación de vanguardia en Machine Learning.

## Data Mining (DM)

### Definición

Técnica para extraer patrones, relaciones o conocimiento útil de grandes volúmenes de datos (usualmente histórico) mediante métodos estadísticos, de aprendizaje automático o heurísticos.

### Aplicación

- Segmentación de Clientes (Clustering): Agrupación de clientes con características similares sin usar etiquetas previas.

**Ejemplo:** Tiendas de retail clasifican clientes para campañas personalizadas.

- Minería de Opiniones / Análisis de Sentimiento: Análisis de comentarios y publicaciones para identificar emociones o percepciones.

**Ejemplo:** Coca-Cola mide en Twitter la aceptación de sus campañas publicitarias.

### Procesamiento

1. Comprensión de los datos: Se define el objetivo del análisis y se explora la información disponible para determinar su relevancia y calidad.
2. Preparación de los datos: Esta etapa incluye varias acciones para asegurar datos de alta calidad:
3. Selección: Se eligen las fuentes de datos y los conjuntos de información más relevantes para los objetivos del proyecto.
4. Limpieza: Se eliminan errores, se completan valores faltantes y se normalizan los datos.
5. Reducción: Se descartan los datos redundantes o de baja calidad que no aportan valor.
6. Transformación: Los datos se adaptan al formato correcto para ser analizados por los algoritmos.

7. Enriquecimiento: Se agregan datos de otras fuentes para aumentar el valor y la calidad de la información.
8. Modelado: Se aplican modelos matemáticos y algoritmos de aprendizaje automático para identificar patrones, asociaciones y correlaciones en los datos limpios y transformados.
9. Evaluación: Se revisan los resultados obtenidos para verificar que el modelo cumple con los objetivos definidos y que los patrones identificados son significativos.

### Batch VS Streaming

- Batch: El Data Mining es inherentemente un proceso de batch, ya que requiere el análisis de grandes volúmenes de datos históricos para descubrir patrones. Los resultados de este análisis no son necesarios en tiempo real.
- Streaming: Aunque el Data Mining tradicional es batch, se están desarrollando técnicas de "Streaming Data Mining" para analizar datos a medida que llegan.

### Herramientas y tecnologías

1. KNIME: Es una plataforma de minería de datos de código abierto muy popular y potente que permite crear flujos de trabajo mediante la conexión de nodos para integrar, manipular, analizar y visualizar datos, además de crear modelos.
2. RapidMiner: Una herramienta versátil para minería de datos, análisis predictivo y aprendizaje automático, que ofrece una amplia gama de funcionalidades para el desarrollo de modelos.
3. Orange: Ofrece visualizaciones de datos atractivas y es especialmente útil para usuarios que requieren crear y explorar modelos sin necesidad de grandes conocimientos técnicos.
4. SAS y SPSS: Son herramientas empresariales que se utilizan para crear modelos de minería de datos, facilitando a los analistas el enfoque en el análisis de patrones en lugar de la codificación.
5. Python: Es un lenguaje ampliamente utilizado para la minería de datos debido a su sintaxis accesible y al rico ecosistema de bibliotecas como Pandas (para manipulación de datos), NumPy (para operaciones numéricas) y Matplotlib (para visualizaciones).
6. R: Es otro lenguaje de programación popular en la comunidad de análisis de datos, utilizado para el análisis estadístico y la creación de modelos de minería de datos.



# Big Data

## Definición

Conjunto de datos tan voluminosos, rápidos (velocidad) y diversos (variedad) que las herramientas tradicionales no pueden procesarlos eficazmente. Se requiere nuevas arquitecturas, escalabilidad y análisis avanzados para extraer valor.

## Aplicación

- **Mantenimiento Predictivo:** Uso de sensores para predecir fallos en máquinas antes de que ocurran.

**Ejemplo:** General Electric aplica Big Data para supervisar turbinas y aerogeneradores.

- **Análisis de Series Temporales y Forecasting:** Predicción de valores futuros basados en datos históricos secuenciales.

**Ejemplo:** Empresas eléctricas predicen la demanda de energía por hora.

## Procesamiento

1. **Captura/Recopilación:** Se obtienen datos de diversas fuentes, como bases de datos empresariales, dispositivos de IoT, redes sociales y sensores.
2. **Almacenamiento:** Los datos se almacenan en plataformas especializadas como Hadoop o bases de datos NoSQL, diseñadas para manejar grandes volúmenes.
3. **Procesamiento:** Se aplican técnicas de procesamiento distribuido para limpiar, transformar y organizar los datos en bruto, preparándolos para el análisis.
4. **Análisis:** Se utilizan herramientas de análisis estadístico, minería de datos y aprendizaje automático (Machine Learning) para extraer patrones, relaciones y tendencias ocultas en los datos.
5. **Visualización:** Los resultados del análisis se presentan de forma gráfica y concisa, usando herramientas y dashboards, para que sean comprensibles y se puedan interpretar fácilmente.

## Batch VS Streaming

- Batch: Históricamente, el Big Data ha estado muy ligado al procesamiento por lotes, utilizando herramientas como MapReduce para procesar grandes conjuntos de datos de una sola vez. Se usa para análisis históricos.
- Streaming: El procesamiento de streaming es cada vez más importante en el Big Data, ya que permite analizar datos en tiempo real a medida que se generan, lo cual es fundamental para aplicaciones como el monitoreo de redes o la detección de fraudes.

## Herramientas y tecnologías

1. Apache Hadoop: Un framework de código abierto para el procesamiento de grandes volúmenes de datos distribuidos, especialmente en lotes.
2. Apache Spark: Un motor de procesamiento rápido para big data que permite realizar análisis en tiempo real y trabajar con lenguaje Scala, según información de The Information Lab.
3. Apache Storm: Una plataforma para el procesamiento de datos en tiempo real y en streaming.
4. MongoDB: Una popular base de datos NoSQL para almacenar y manejar datos no estructurados.
5. Apache Cassandra: Una base de datos distribuida de alta disponibilidad diseñada para manejar grandes volúmenes de datos y asegurar tolerancia a fallos.
6. Elasticsearch: Un motor de búsqueda y analítica distribuido, basado en documentos JSON, para indexar y analizar grandes volúmenes de datos en tiempo real.
7. Python: Un lenguaje versátil y popular para big data, especialmente por sus librerías para análisis, según UDIT y The Information Lab Spain.
8. R: Un lenguaje de código abierto centrado en el análisis estadístico de los datos.
9. Scala: El lenguaje principal para el desarrollo de aplicaciones en Apache Spark.
10. Tableau y Power BI: Herramientas para la visualización y análisis de datos que permiten crear dashboards interactivos para los tomadores de decisiones.
11. Kibana: Una herramienta para visualizar y explorar datos en tiempo real, utilizando dashboards interactivos sobre datos almacenados en Elasticsearch.

12. Logstash: Parte del stack ELK (Elasticsearch, Logstash, Kibana), esta herramienta se encarga de extraer datos de diversas fuentes, transformarlos y almacenarlos.

## Conclusión

El avance de la Inteligencia Artificial, el Machine Learning, la Minería de Datos y el Big Data demuestra cómo la tecnología puede transformar la manera en que vivimos y trabajamos. Si bien ofrecen muchos beneficios al facilitar tareas y mejorar la toma de decisiones, también nos invitan a reflexionar sobre el uso responsable de los datos y el impacto que tienen en la sociedad. Entender estas herramientas no solo es útil para aprovecharlas, sino también para usarlas de forma ética y consciente en nuestro día a día.

### Big Data

La materia prima. Grandes conjuntos de datos que se procesan.

### Data Mining

El proceso para encontrar patrones y conocimiento en el Big Data.

### Inteligencia Artificial (IA)

El campo más amplio. El objetivo final: crear inteligencia.

### Machine Learning (ML)

Un subcampo de la IA. Permite que las máquinas aprendan de los datos.

## Referencias Bibliográficas

*¿Qué es la inteligencia artificial o IA? | Google Cloud.* (n.d.). Google Cloud.  
<https://cloud.google.com/learn/what-is-artificial-intelligence?hl=es-419>

*Blog: Fundamentos del Machine Learning: Guía Completa para Iniciar en el Aprendizaje Automático.* (n.d.). <https://www.kranio.io/blog/fundamentos-del-machine-learning>

Guru. (2025, July 15). *Herramientas de IA: Una guía completa.*  
<https://www.getguru.com/es/reference/ai-tools>

*Herramienta de minería de datos: descripción general | Temas de ScienceDirect.* (n.d.).  
<https://www.sciencedirect.com/topics/computer-science/data-mining-tool>

Ramírez, L. (2024, October 29). *Las 10 mejores herramientas de Big Data 2024.* Thinking for Innovation. <https://www.iebschool.com/hub/mejores-herramientas-big-data/>

Ortega, C. (2023, October 1). *Herramientas de inteligencia artificial: 5 ejemplos y sus características.* QuestionPro. <https://www.questionpro.com/blog/es/herramientas-de-inteligencia-artificial/>

Universidades, S. (2025, April 22). *Data mining: qué es, técnicas y ejemplos | Blog Santander Open Academy.* Santander Open Academy.  
<https://www.santanderopenacademy.com/es/blog/data-mining-que-es.html>