



**AWAKELAB**

**BASECAMP**

Ciencia de Datos

## Módulo: Aprendizaje de Máquina Supervisado

---

### Aprendizaje Esperado

---

1. Describir los conceptos fundamentales de aprendizaje de máquina para resolver un problema.

---

### Aprendizaje de máquina

#### Fundamentos

El aprendizaje de máquina es eso, una máquina aprendiendo sobre los datos, es también conocido como Machine Learning y tiene como objetivo identificar patrones a partir de los datos con el fin de hacer predicciones, detecciones o clasificaciones.

#### Origen del Machine Learning

- Arthur Samuel (1901-1990) propone en 1959 darle a las máquinas y computadoras la habilidad de aprender sin ser explícitamente programadas.
- El informático fue pionero en el área de la Inteligencia Artificial y en comprender que observar patrones ayudaba a replicarlos.
- Gracias al desarrollo teórico de estadísticos e informáticos, el Machine Learning se hizo muy famoso en los años 90.
- La intersección de la informática y las estadísticas dio lugar a un enfoque probabilístico en la Inteligencia Artificial.

#### ¿Qué es el aprendizaje de máquina?

El aprendizaje de máquina es una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos para imitar la forma en que los humanos aprenden, mejorando gradualmente su precisión.

El aprendizaje automático es un componente importante del creciente campo de la ciencia de datos. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones, descubriendo información clave dentro de los proyectos de minería de datos. Estos conocimientos posteriormente impulsan la toma de decisiones dentro de las aplicaciones y los negocios, lo que idealmente impacta en las métricas de crecimiento clave. A medida que el big data continúe expandiéndose y creciendo, aumentará la demanda del mercado de científicos de datos, lo que requerirá que ayuden a identificar las preguntas comerciales más relevantes y, posteriormente, los datos para responderlas.

## Aplicaciones

Estos son solo algunos ejemplos de aprendizaje automático que puede encontrar todos los días:

Área	Ejemplo
Reconocimiento de imágenes	Detección de personas sin mascarilla.
Clasificación	Análisis de clientes para evaluación de entrega de crédito bancario.
Reconocimiento de voz	Atención al cliente mediante llamadas telefónicas.
Optimización	Determinar rutas óptimas en tiempo real.
Reconocimiento de imágenes	Detección de espacios vacíos en góndolas de supermercados.
Predicción	Detección de vuelos con alta probabilidad de atraso.
Optimización	Determinar dotación de camiones óptima según demanda minera.

Entre otras diversas aplicaciones.

## Etapas de un Aprendizaje de Máquina

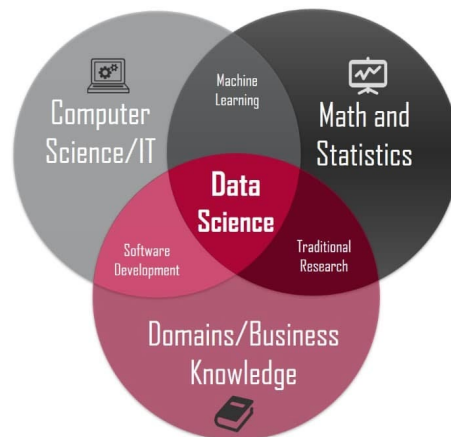
### ¿Qué necesitamos para resolver problemas analíticos?

Para resolver cualquiera de los ejemplos anteriores, se necesitan conocimientos de:

- **Big Data:** Volumen masivo de datos (las cinco V del Big data son Volumen, Variedad, Velocidad, Veracidad y Valor).

- **Data Mining:** Extracción de información a partir de los datos.
- **Business Intelligence:** Análisis descriptivo, muestra lo que sucede/sucedió, visualización y comunicación adecuada de los datos en el contexto del negocio.
- **Machine Learning:** Algoritmos que acceden y aprenden de los los datos.
- **Deep Learning:** Algoritmos que buscan comprender estructuras abstractas y complejas de los datos.
- **Business Analytics:** Técnicas predictivas, determina la probabilidad de resultados futuros con interpretación en el negocio.

En esta parte la **ciencia de datos** es una combinación de varias herramientas, algoritmos y principios, conformando un estudio interdisciplinar de los datos con el fin de generar conocimiento.



### Etapas típicas de un problema de aprendizaje de máquina

- Hacer las preguntas correctas, determinar un dolor, una necesidad del negocio.
- Obtener y recopilar datos, estructurar los datos limpios en un formato apropiado y realizando reducción de la dimensionalidad, estandarización, normalización, entre otras.
- Realizar gráficos, análisis exploratorio, generar intuiciones de negocio. Determinar los features a utilizar, crear o transformar variables.

- Modelar el algoritmo de Machine Learning, se construye con los datos de entrenamiento. Evaluar el modelo obtenido y ajustarlo para maximizar su rendimiento.
- A partir del modelo creado, obtener inferencias e insights que agreguen valor de negocio.

## Herramientas útiles para un proyecto de Data Science

- **Data Mining**



R es un lenguaje de programación vastamente utilizado a la hora de realizar minería de datos, análisis descriptivo, limpieza, incluso creación de gráficas, reportes automáticos y dashboards.

Python es un lenguaje de programación que destaca por su versatilidad y fácil integración con otras aplicaciones y plataformas.



- **Visualización: Business Intelligence**

Power BI y Tableau son herramientas que permiten crear informes con múltiples visualizaciones interactivas mediante una interfaz sencilla de utilizar.



### **Tipos de algoritmos**

#### **Aprendizaje supervisado**

También conocido como aprendizaje automático supervisado, se define por el uso de conjuntos de datos etiquetados para entrenar algoritmos



que clasifican datos o predicen resultados con precisión. A medida que los datos de entrada se introducen en el modelo, éste ajusta sus pesos hasta que el modelo se haya ajustado correctamente. Esto ocurre como parte del proceso de validación cruzada para garantizar que el modelo evite el sobreajuste o el ajuste insuficiente. El aprendizaje supervisado ayuda a las organizaciones a resolver una variedad de problemas del mundo real a escala, como clasificar el correo no deseado en una carpeta separada de su bandeja de entrada.

Un algoritmo tiene un aprendizaje supervisado cuando se busca predecir casos futuros a partir de datos cuya respuesta *ya se conoce*. El algoritmo aprende a partir de las variables explicativas asociadas a la variable respuesta (numérica o categórica). Luego, predice el valor de la variable respuesta cuando se presentan nuevas variables explicativas.

Algunos métodos utilizados en el aprendizaje supervisado incluyen *redes neuronales, bayes ingenuos, regresión lineal, regresión logística, bosque aleatorio, máquina de vectores de soporte (SVM) y más*.

### **Aprendizaje no supervisado**

Conocido también como aprendizaje automático no supervisado, utiliza algoritmos de aprendizaje automático para analizar y agrupar conjuntos de datos no etiquetados. Estos algoritmos descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana. Su capacidad para descubrir similitudes y diferencias en la información lo convierte en la solución ideal para el análisis exploratorio de datos, estrategias de venta cruzada, segmentación de clientes, reconocimiento de imágenes y patrones.

El algoritmo aprende a partir de la *variable explicativa sin ninguna variable respuesta asociada*, lo que le permite determinar los patrones de datos por sí mismo. Los algoritmos se dejan guiar por sus propios mecanismos para descubrir la estructura de datos.

También se usa para reducir la cantidad de características en un modelo a través del proceso de reducción de dimensionalidad; El análisis de componentes principales (PCA) y la descomposición de valores singulares (SVD) son dos enfoques comunes para esto. Otros algoritmos utilizados en el aprendizaje no supervisado incluyen redes neuronales, agrupamiento de k-means, métodos de agrupamiento probabilístico y más.

## **Clasificación vs Regresión**

Los problemas de modelamiento predictivo de clasificación son diferentes a los problemas de modelamiento predictivo de regresión.

- La clasificación es la tarea de predecir una etiqueta de clase discreta.
- La regresión es la tarea de predecir una cantidad continua.

Existe cierta superposición entre los algoritmos de clasificación y regresión; por ejemplo:

- Un algoritmo de clasificación puede predecir un valor continuo, pero el valor continuo tiene la forma de una probabilidad para una etiqueta de clase.
- Un algoritmo de regresión puede predecir un valor discreto, pero el valor discreto en forma de cantidad entera.

Algunos algoritmos se pueden usar tanto para clasificación como para regresión con pequeñas modificaciones, como árboles de decisión y redes neuronales artificiales. Algunos algoritmos no pueden, o no pueden usarse fácilmente para ambos tipos de problemas, como la regresión lineal para el modelado predictivo de regresión y la regresión logística para el modelado predictivo de clasificación.

Es importante destacar que la forma en que evaluamos las predicciones de clasificación y regresión varía y no se superpone, por ejemplo: Las predicciones de clasificación se pueden evaluar utilizando la precisión, mientras que las predicciones de regresión no. Las predicciones de regresión se pueden evaluar utilizando el error cuadrático medio, mientras que las predicciones de clasificación no. Estudiaremos en detalle esto más adelante.



## Referencias

[1] Aprendizaje de Máquina

<https://www.ibm.com/cloud/learn/machine-learning>

[2] Machine learning

<https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>

[3] Video: Introducción al Aprendizaje Automático en la ESA

<https://www.youtube.com/watch?v=FXYPDyoR1o0>

[4] ¿Clasificación o regresión?

<https://www.iartificial.net/clasificacion-o-regresion/>

## Material Complementario

[1] Aprendizaje de máquina o Machine learning

<https://www.youtube.com/watch?v=NdNyYcAJQr8>

[2] ¿Qué es el aprendiza de máquina?

<https://www.youtube.com/watch?v=gDsARHubwMo>

[3] Aprendizaje supervisado y no supervisado

<https://www.youtube.com/watch?v=oT3arRRB2Cw>