



SKILLING
CENTER

TECMILENIO



Inteligencia Artificial y Machine Learning

Introducción al machine
learning





La inteligencia artificial está presente e integrada en todas las actividades diarias. Como consumidores, es evidente que cada vez más productos incorporan la inteligencia artificial en su propuesta de valor.

En la industria financiera, estos productos y servicios inteligentes tienen un objetivo explícito, y su labor es encontrar la configuración que optimice este objetivo. Por ejemplo, es posible modificar las condiciones de oferta para una solicitud crediticia ajustando las condiciones de la oferta, dado el perfil de riesgo y comportamiento de pagos del solicitante.

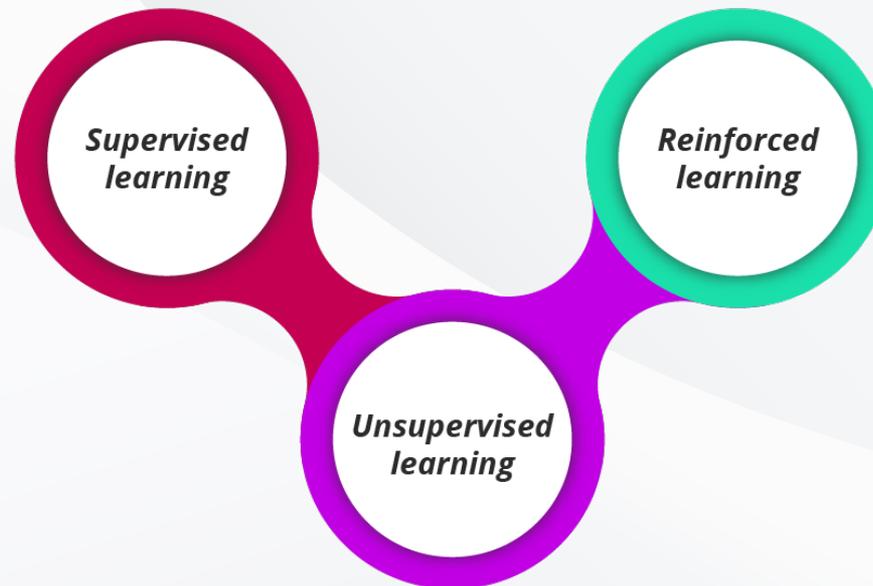
En esta experiencia educativa comprenderás lo que se puede lograr con esta tecnología y la forma de aplicarlo para generar valor en la toma de decisiones.





La inteligencia artificial es una disciplina que lleva años en desarrollo. En la actualidad, se ha visto una mejora significativa en los resultados que se pueden obtener de dicha disciplina.

Existen varios enfoques o formas en que las máquinas aprenden en la práctica de *machine learning*. Considerando que utilizan algoritmos (conjunto de operaciones sistémicas) para resolver problemas, Brown (2021) utiliza tres subcategorías representativas del aprendizaje automático:



Hablar de datos etiquetados se refiere a que los datos que usará el algoritmo para solucionar el problema que han sido identificados correctamente antes de ser procesados.

En la opinión de Malone et al. (2020), las técnicas recomendadas en la creación de sistemas computacionales para detectar patrones, o bien, tomar acciones con base en datos recabados son las siguientes:





Una base de datos es un conjunto de información estructurada o *dataset* que pertenece al mismo contexto, agrupada con el fin de administrar la información. Cuando se habla de una base de datos, se tiene que diferenciar entre sus dos componentes principales:

Variables independientes o *features*

Variables dependientes u objetivo

Los algoritmos de aprendizaje supervisado requieren de una variable dependiente para el proceso de entrenamiento del algoritmo.



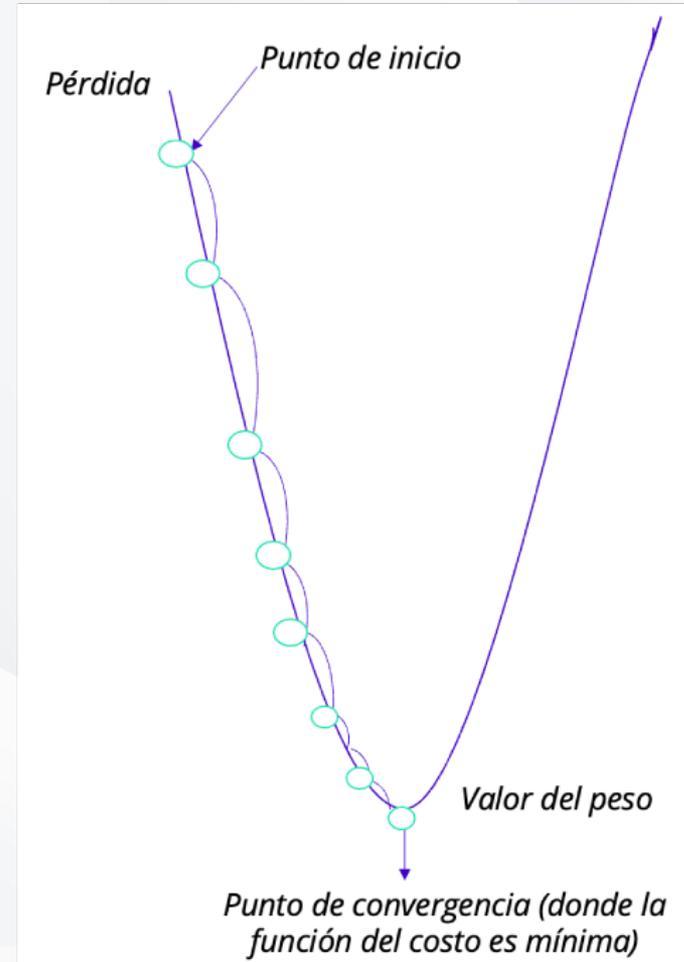
¿Por qué son tan importantes los datos en los algoritmos de aprendizaje? La respuesta es simple. Los algoritmos supervisados dependen de la información que tengan disponible para encontrar una relación entre las variables descriptoras y las variables objetivo. Esto convierte al algoritmo supervisado en una función:

$$f: x \rightarrow y$$



El proceso de “entrenamiento” le permite al modelo “aprender” a generar la relación entre x y y . De acuerdo con Huyen (2022), el proceso de aprendizaje requiere que se defina una función objetivo que se tiene que maximizar (función de ganancia) o minimizar (función de error).

Uno de los algoritmos más utilizados para optimizar la función objetivo de forma automática es el gradiente descendente o gradiente ascendente en caso de tener una función de ganancia.





A fin de comprender las funciones y utilidad del machine learning y los procesos que conlleva su desarrollo, realiza lo siguiente:

1. Investiga algún modelo de evaluación de riesgo que se haya generado.
2. Descríbelo y evalúalo de forma crítica.
3. Encuentra soluciones que beneficien el cumplimiento del objetivo (riesgo) por el cual se creó, detállalo.



Entendiendo el contexto y los conceptos fundamentales del machine learning, estás en una mejor posición para comenzar a implementar en la práctica diferentes tipos de modelos de aprendizaje automático.

Lo aprendido te permitirá identificar las diferentes subcategorías del machine learning y cuándo es conveniente usar una sobre otra. Esto es sumamente importante en la práctica, dado que los algoritmos en cada subcategoría se especializan en resolver un tipo de problema en particular de la forma más eficiente.

Desde el enfoque de algoritmos de aprendizaje supervisado, se identifican dos tipos de tareas principales: algoritmos de regresión y clasificación. Con esta base, tienes la capacidad para distinguir cuándo es adecuado usar uno sobre otro y entender los fundamentos algorítmicos del proceso de aprendizaje.



- Brown, S. (2021). *Machine learning, explained*. Recuperado de <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>
- Huyen, C. (2022). *Designing machine learning system*. O'Reilly Media, Inc.
- Malone, T., Rus, D., y Laubacher, R. (2020). *Artificial intelligence and the future of work*. Recuperado de <https://workofthefuture.mit.edu/research-post/artificial-intelligence-and-the-future-of-work/>



SKILLING
CENTER

TECMILENIO



Inteligencia Artificial y Machine Learning

Análisis de datos





Una de las funciones de las instituciones financieras es sin duda la aprobación o el rechazo de solicitudes de préstamos y servicios, para lo que utilizan, en algunos casos, modelos estadísticos en una base de datos con el fin de establecer el nivel de riesgo del solicitante, que puede ser alto, medio o bajo, y tener bases para aprobar o rechazar la solicitud. El modelo cuenta con un algoritmo para clasificar el riesgo del solicitante, pero ¿qué hay detrás de ese algoritmo que a un experto en finanzas que está aprendiendo analítica avanzada le debe interesar?

En esta experiencia educativa aprenderás lo relacionado a la estructura, preparación y limpieza de la información que alimentará los modelos diseñados para apoyar en la toma de decisiones.





En la industria fintech, gran parte de la información es almacenada en manejadores de base de datos o sistemas de gestión de base de datos (SGBD o DBMS, por sus siglas en inglés). La generación de estos datos proviene principalmente de la ejecución de una serie de algoritmos que representan un modelo de negocio.





En la práctica, el ambiente digital que contiene la lógica de negocio es conocido como OLTP (*Online Transaction Processing*). Según Sinha (2021), este ambiente es el principal generador de la información, ya que contiene la operación de la lógica de negocio. Por otra parte, el ambiente analítico, conocido como OLAP (*Online Analytical Processing*) se enfoca en proveer análisis de alto valor utilizando los datos generados por la operación.

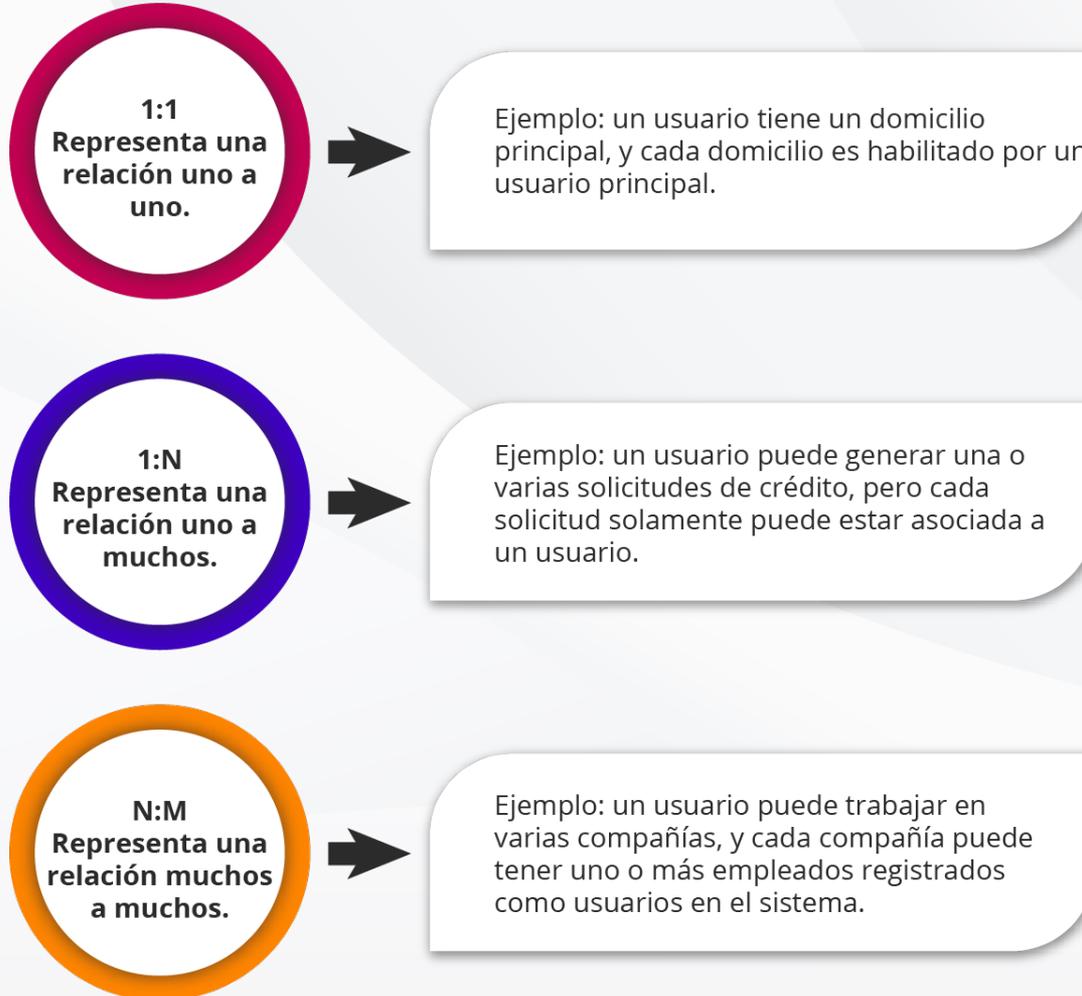


Es común que el ambiente OLTP represente a las entidades de negocio con programación orientada a objetos. Estas entidades contribuyen significativamente en el diseño de las bases de datos. En el caso de la solicitud crediticia, se encuentran dos entidades de negocio fáciles de distinguir: la entidad de una solicitud y la entidad de usuario.

Una vez identificada la entidad en el proceso de negocio se puede crear una tabla en la base de datos para almacenar las distintas instancias de la entidad que se generen en la operación. Siguiendo el ejemplo, esto significa que se puede crear una tabla de usuarios y otra de solicitudes.



Existen diferentes tipos de relaciones o cardinalidades:





Si se considera como ejemplo, el usuario y la solicitud crediticia recibidas por la institución de crédito y almacenadas en una base de datos relacional, la cardinalidad entre usuarios y solicitudes es uno-a-muchos, ya que un usuario puede tener varias solicitudes, pero una solicitud solo corresponde a un usuario en particular.

Al momento de hacer una consulta que involucre ambas tablas, se tendrá como resultado un conjunto de datos en donde ninguna solicitud sea repetida, pero los usuarios aparecen potencialmente más de una vez.

Usando la siguiente figura como referencia, se puede apreciar cómo el conjunto de datos resultante del producto entre la tabla user y credit_request tiene una ocurrencia del usuario 1 y dos ocurrencias del usuario 2.

Combinando tablas de datos - explicación de Joins de SQL

Una cláusula de Join en SQL es usada para combinar renglones de dos o más tablas, basado en una columna o campo que las relaciona.

Tabla 1 ●

1		
2		

Tabla 2 ●

1		
3		
4		

Outer Join ●●

1				
2				
3				
4				

Union ●+●

1		
2		
1		
3		
4		

Inner Join ●●

1				

Left Join ●●

1				
2				

Cross Join ●⊗

1			1		
1			3		
1			4		
2			1		
2			3		
2			4		



Según Arya (2022), modelar la lógica de negocios utilizando entidades y tablas relacionales permite ser eficientes en el procesamiento atómico de las transacciones de negocio, lo cual es un incentivo fuerte para adoptarlo en la operación del sistema OLTP. Por el contrario, desnormalizar las tablas relacionales permite crear conjuntos de datos en donde el tamaño del *dataset* se rige por una entidad en particular y contiene a su vez la información de otras entidades relacionadas.



El proceso de manipulación y preparación de datos para su limpieza y ordenamiento es una de las labores más intensivas en la ciencia de datos.



Obtener el estado final de los datos requiere de un proceso iterativo de exploración y visualización de los mismos. De acuerdo con Ghosh (2022), es posible sintetizarlos de la siguiente manera:

1. Tratamiento de valores ausentes y valores atípicos

2. Estandarización de variables continuas

3. Transformaciones adicionales (opcional)

4. Categorización de variables continuas (opcional)

5. Codificación de variables categóricas (opcional)



Como asesor financiero es necesario considerar bases de datos efectivas, por lo que se te pide realizar lo siguiente:

1. Construye una base de datos de al menos 25 participantes.
2. Identifica la cantidad de variables que pueden formarla, por ejemplo, nombre, dirección, teléfono, tipos de servicios solicitados, capacidad de pago, etc. (justifica las variables requeridas).
3. Establece la relación de las variables (dependiente e independientes).
4. Identifica y comenta el modelo a trabajar con ella, es decir, ¿cuál sería el modelo predictivo para una toma de decisiones efectiva? Descríbelo.



La utilidad del modelo para generar el nivel de riesgo de los clientes, y la forma en la que el modelo puede ser creado, tiene más herramientas para utilizar en la toma de decisiones, no solo para el nivel de riesgo sino para evaluar el tipo de rentabilidad a obtener, el máximo límite a otorgar, tiempo del crédito, etcétera.

En la práctica, la información que se necesita para desarrollar modelos de aprendizaje automático está esparcida a través de varios sistemas. En la industria es común que el mayor generador de la información sea la operación del sistema transaccional (OLTP), dado que contiene la representación de la lógica de negocio.

La generación de un conjunto de datos que sirva para crear modelos de machine learning depende entonces de la capacidad de manipular, integrar, limpiar y reconciliar las distintas fuentes. Los primeros tres pasos del proceso son sumamente importantes:

1. Obtención de la información en crudo.
2. Ordenamiento de los datos (*tidy dataset*).
3. Transformaciones de limpieza.

Dominando estos pasos, tendrás la capacidad de crear datasets de alta calidad para algoritmos de machine learning.



- Arya, S. (2022). *Entity in DMBS*. Recuperado de <https://www.scaler.com/topics/entity-in-dbms/>
- Ghosh, S. (2022). *A comprehensive guide de data preprocessing*. Recuperado de <https://neptune.ai/blog/data-preprocessing-guide>
- Sinha, T. (2021). *OLAP vs. OLTP: What's the difference?* Recuperado de <https://www.ibm.com/cloud/blog/olap-vs-oltp>

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.

Todos los derechos reservados @ Universidad Tecmilenio

La obra presentada es propiedad de ENSEÑANZA E INVESTIGACIÓN SUPERIOR A.C. (UNIVERSIDAD TECMILENIO), protegida por la Ley Federal de Derecho de Autor; la alteración o deformación de una obra, así como su reproducción, exhibición o ejecución pública sin el consentimiento de su autor y titular de los derechos correspondientes es constitutivo de un delito tipificado en la Ley Federal de Derechos de Autor, así como en las Leyes Internacionales de Derecho de Autor. El uso de imágenes, fragmentos de videos, fragmentos de eventos culturales, programas y demás material que sea objeto de protección de los derechos de autor, es exclusivamente para fines educativos e informativos, y cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por UNIVERSIDAD TECMILENIO. Queda prohibido copiar, reproducir, distribuir, publicar, transmitir, difundir, o en cualquier modo explotar cualquier parte de esta obra sin la autorización previa por escrito de UNIVERSIDAD TECMILENIO. Sin embargo, usted podrá bajar material a su computadora personal para uso exclusivamente personal o educacional y no comercial limitado a una copia por página. No se podrá remover o alterar de la copia ninguna leyenda de Derechos de Autor o la que manifieste la autoría del material.