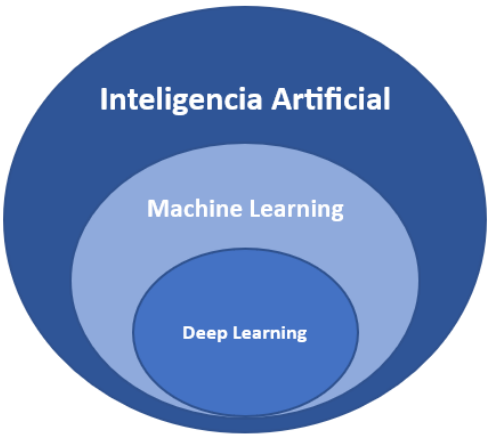
Introducción al Aprendizaje Automático

Contexto dentro de la Inteligencia Artificial

Los algoritmos de machine learning (ML) forman parte de lo se conoce como Inteligencia Artificial (IA).

Las características de la IA se pueden resumir como computadores con la habilidad de razonar como los humanos, los algoritmos de ML agregan la capacidad de aprender desde los datos sin necesidad de ser programados y los algoritmos de Aprendizaje Profundo (Deep Learning) tienen la capacidad de adaptarse a nuevos datos.



Relación IA, ML, DL

Definición

Tom Mitchell definió al aprendizaje automático (machine learning) como " *el estudio de algoritmos de computación que mejoran automáticamente su rendimiento gracias a la experiencia*".
Más formalmente lo define como: " *Se dice que un programa de computadora aprende de la experiencia E con respecto a una clase de tareas T y medida de desempeño P, si su desempeño en tareas T medidas por P mejora con la experiencia E*."
En definitiva, son algoritmos tienen la capacidad de resolver una determinada tarea sin necesidad de ser programados para tal fin, sino que aprenden a resolverlas a partir de ser entrenados con un conjunto de datos. Y mejoran su capacidad para resolver esa tarea a partir de la experiencia.
La selección del algoritmo dependerá de la tarea a realizar y los datos deberán ser preprocesados para adaptarse a los tipos de entrada que espera el algoritmo.
Estos algoritmos identifican patrones en los datos a partir de la utilización de métodos estadísticos y en base a los mismos pueden realizar una tarea específica. Tal como se indica en su definición, para poder determinar si aprenden es necesario medir su perfomance a partir de diferentes métricas.

El modelo

Un modelo es una construcción simplificada de una realidad más compleja que se utiliza para comprenderla.
Para la construcción del modelo se seleccionarán una serie de características (variables) de la realidad.
Cuanto más variables, más cercana a la realidad será la representación pero también más compleja, por lo tanto la selección de las características no es una tarea menor importancia.

Tipos de tareas

Los sistemas de ML son capaces de procesar grandes volúmenes de datos, identifican patrones de comportamiento y a partir de ellos predicen comportamientos futuros.
Abordan diferentes tipos de tareas como pueden ser:

- Clasificación: el modelo es capaz de asociar una instancia a una clase.
- Regresión: el modelo es capaz de predecir el valor que tendrá la variable de salida.
- Agrupamiento o Clusterización: el modelo es capaz de realizar agrupamiento de los datos a partir de la identificación automática de patrones.
- Identificar asociaciones: el modelo es capaz de identificar reglas de asociación a partir de los datos.
- Reducir: el modelo es capaz de encontrar dependencias entre los atributos de las instancias y reducir la dimensionalidad de la instancia.

El proceso

El aprendizaje automático es un proceso iterativo que requiere que cada paso sea revisado a medida que se aprende más sobre el problema bajo investigación. Este proceso iterativo puede requerir el uso de diferentes herramientas para cada paso.

Proceso de ML

Definir el problema: investigar y caracterizar el problema para entender mejor los objetivos que deberá alcanzar el modelo

Analizar los datos: realizar un análisis exploratorio para describir estadísticamente y visualizar los datos para los datos que se tienen disponibles

Seleccionar el algoritmo: dependiendo del problema a resolver, es decir el tipo de tarea que deberá realizar el modelo, y de los datos que se tienen disponibles se deberá seleccionar el algoritmo que resuelva dicha tarea.

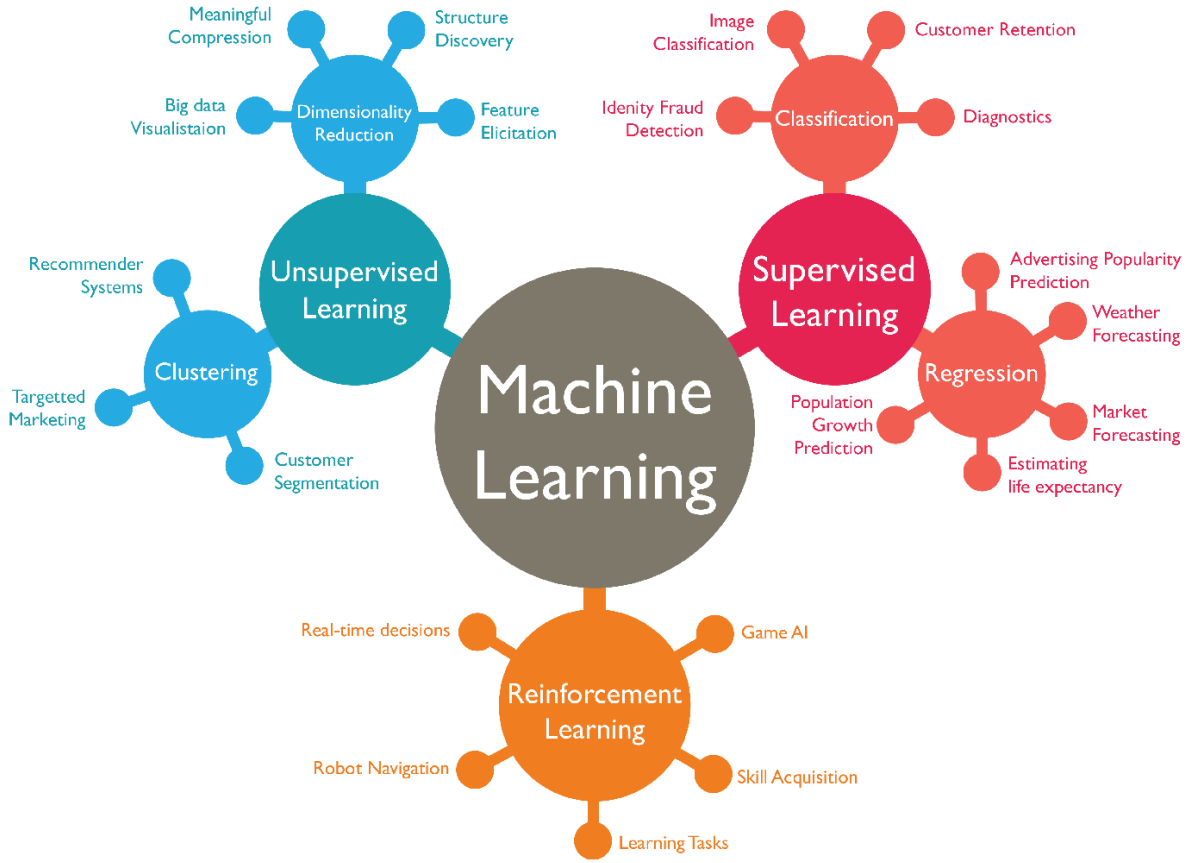
Preparar los datos: se exploran los datos y se realizan tareas de limpieza y transformación para que se adecuen a las estructuras de datos de entrada que espera el algoritmo.

Entrenar, Evaluar y Mejorar el modelo: se ejecuta el modelo con un lote de datos con el objetivo de que el algoritmo aprenda de los mismos. Se mide la calidad del modelo y se realizan ajustes, incluso se pueden aplicar modelos ensamblados para obtener mejores resultados. En este punto se puede volver a cualquiera de las etapas anteriores del proceso.

Presentar resultados: desplegar el modelo para que realice predicciones y presente los resultados.

Tipos de Aprendizaje Automático

Tal como dijimos en secciones anteriores los algoritmos de ML aprenden de los datos a partir de una etapa de entrenamiento.
Existen distintos tipos de entrenamiento: supervisado, no supervisado y por refuerzo.



Clasificación del ML según entrenamiento y tarea

Supervisado: cada instancia de los datos de entrenamiento tendrá una “etiqueta” con el resultado esperado para esa instancia.

No Supervisado: no requieren que los datos de entrenamiento posean la etiqueta de resultado esperado, sino que identifican patrones y devuelven los tipos encontrados según sus similitudes.

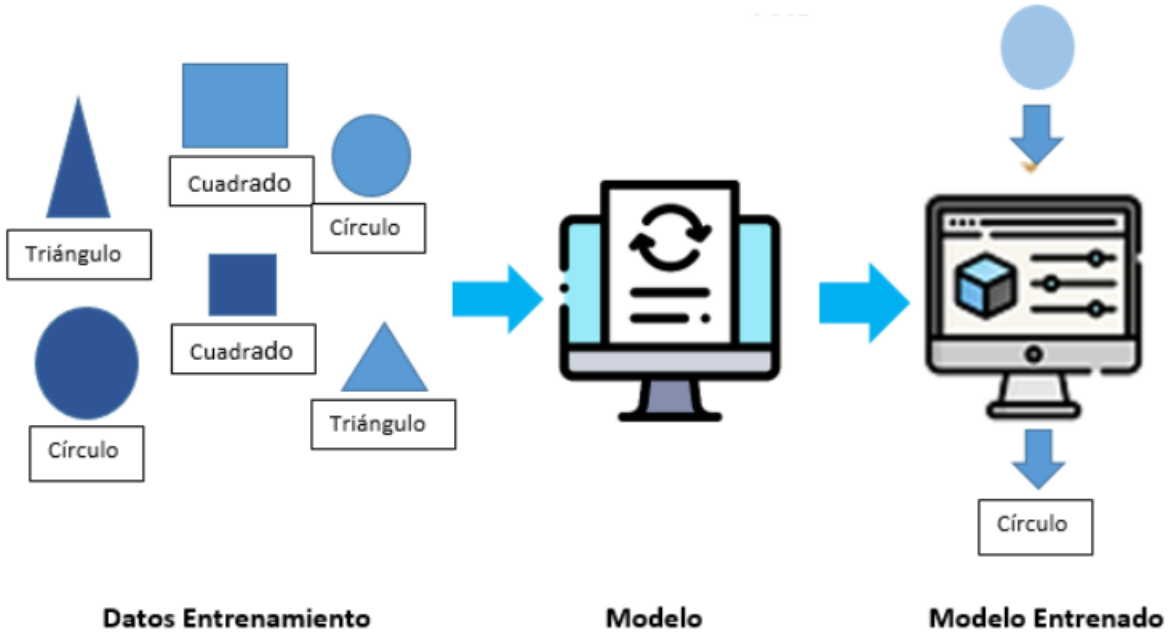
Por refuerzo: son modelos que poseen un agente que interactúa con un entorno y el aprendizaje consiste en la recepción de recompensas del entorno .No serán vistos en esta materia.

Aprendizaje Automático Supervisado

Es el uso de datos de entrenamiento etiquetados con los resultados esperados. Es decir, se necesita la intervención humana para la preparación de los datos de entrenamiento.

A medida que los datos de entrenamiento se introducen en el modelo, este ajusta sus ponderaciones hasta que dicho modelo se haya ajustado adecuadamente.

Este conjunto de datos incluye datos de entrada y resultados correctos, que permiten que el modelo aprenda con el tiempo. El algoritmo mide su precisión a través de la función de pérdida, ajustándose hasta que el error se haya minimizado lo suficiente.



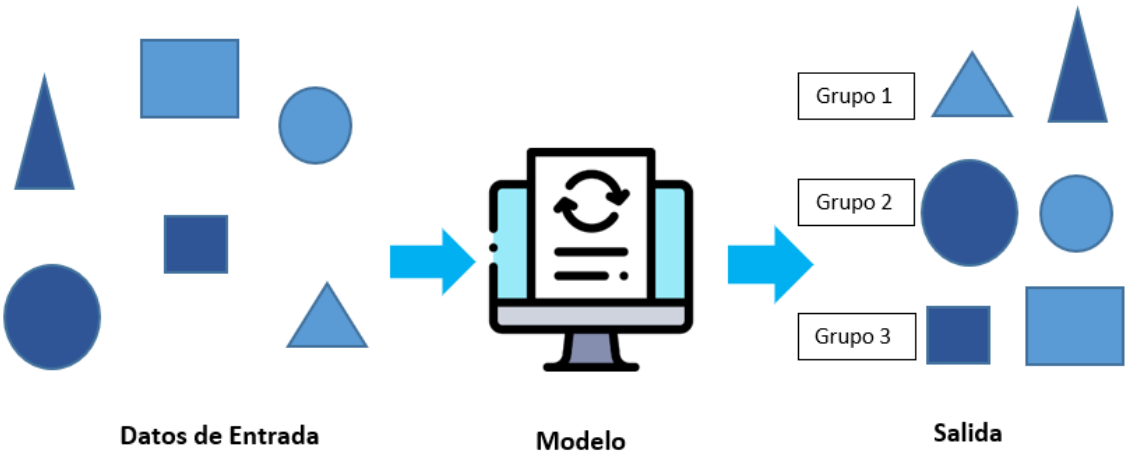
Aprendizaje Automático Supervisado

Aprendizaje Automático No Supervisado

Utiliza algoritmos que son capaces de analizar los datos de entrada e identificar patrones y características comunes que les permiten agruparlos. Es decir, estos algoritmos analizan los datos sin intervención humana para su etiquetado, son capaces de descubrir similitudes y diferencias en la información.

Resultan muy útiles para análisis de datos exploratorios identificando dependencias entre las variables de las instancias agrupandolas y son utilizados para solucionar el problema de la multidimensionalidad. Muchos algoritmos no tienen buena calidad cuando las instancias de entrada tienen gran cantidad de variables por lo tanto es necesario reducirlas, en esos casos estos algoritmos resultan de suma utilidad.

Otro ejemplo de problemas que resuelven es el agrupamiento, también conocido como clusterización, los cuales tienen múltiples aplicaciones como pueden ser la segmentación de clientes.



Aprendizaje Automático No Supervisado

Preprocesamiento

El preprocesamiento es el tratamiento que debe darse a los datos previamente al entrenamiento del algoritmo. Podríamos incluir en esta denominación tanto al análisis exploratorio de los datos con los que contamos como también la preparación de dichos datos para que se ajusten a la forma de entrada que necesita el algoritmo seleccionado.
Los datos tal cual los obtenemos de sus fuentes generalmente no se presentan de la forma que lo necesitan los algoritmos de IA.
Por ejemplo hay modelos que como datos de entrada sólo aceptan variables numéricas, por lo tanto si nuestros datos tienen variables categóricas que consideramos que pueden ser buenos predictores deberemos pre procesarlas convirtiéndolas en numéricas con una técnica que se conoce como “variables dummy”.
Otros algoritmos son sensibles a los valores outliers y por lo tanto hay que suprimirlos o “suavizarlos”, del análisis exploratorio también pueden surgir variables con gran cantidad de valores nulos los cuales deberán completarse o suprimirse.
Y así, múltiples problemas pueden presentarse que deben solucionarse con las tareas de preprocesamiento, las cuales consisten en la aplicación de técnicas con el objetivo de adecuar los datos para ser utilizados
Tal como vimos anteriormente, el proceso de la construcción y puesta en producción de un modelo de machine learning no es lineal. En cualquier etapa del proceso puede volverse a una etapa anterior. Lo mismo ocurre con el preprocesamiento.
Se considera que el 70% del tiempo de la construcción de un modelo corresponde a tareas de preprocesamiento y son tareas críticas para la calidad del modelo.

Recolección e integración de los datos

La primera dificultad que tenemos que sortear una vez que tenemos identificado y reconocido el problema de negocio que deberá resolver nuestro modelo es la obtención de los datos que describen dicho negocio en el marco del problema a resolver.

Generalmente debe recurrirse a diversas fuentes tanto internas como externas a la organización. Hay que analizar la calidad de esas fuentes, su disponibilidad, y recurrencia de actualización.
No siempre se cuenta con fuentes de estas características y por lo tanto se debe re-definir el problema a resolver de acuerdo a las fuentes de datos con las que contamos.
Una vez obtenidas, el análisis de datos puede requerir una combinación de datos de múltiples formatos (MS Excel, BD relacionales, data warehouse) y se los debe convertir en un conjunto de datos coherentes entre sí.
Debido a que provienen de diferentes fuentes pueden presentarse diferentes problemas de inconsistencias, ambigüedades o diferencias para un mismo dato como por ejemplo:

- diferencias en las denominaciones (Client_id, IdCliente),
- heterogeneidad semántica (alto- medio- bajo en diversas fuentes significan lo mismo?)
- conflictos de valores de datos (diferentes escalas, diferentes representaciones)
- registros redundantes
- atributos redundantes

Análisis exploratorio de los datos

Una vez obtenidos los datos desde las diversas fuentes es necesaria la exploración de los mismos para identificar valores inusuales, valores extremos, valores nulos, discontinuidades y otras particularidades de los mismos.
Se deben usar los datos estadísticos y visualizarlos de manera gráfica. Esto nos permitirá comprender aún más el comportamiento del negocio, descubrir relaciones no indagadas hasta el momento, evaluar su calidad e incluso podría condicionar la selección del algoritmo a utilizar en nuestro modelo.

Limpieza de datos

Una vez recolectados los datos muy probablemente no puedan ser utilizados tal cual los obtuvimos ya que pueden presentar diversos problemas como pueden ser:

- Ruido
- Registros duplicados
- Datos incompletos
- Datos inconsistentes
- Formato inadecuado
- Grandes volúmenes

La limpieza de los datos son tareas tendientes a completar los datos faltantes, suavizar los datos con ruido, corregir inconsistencias y eliminar los valores outliers (datos atípicos)

Transformación de datos

Puede ocurrir que los datos no tengan la forma o tipo que el algoritmo requiere como entrada y por lo tanto se deberán realizar transformaciones sobre los mismos para que se adecuen a la estructura y forma esperada.
Como ejemplos de tareas de transformación se pueden mencionar:

- Normalización: reducir la redundancia y ambigüedad, utilizar un mismo valor para una misma denominación, simplificadamente comparar peras con peras
- Suavizado: elimina el ruido permitiendo que se destaquen patrones relevantes

- Agregación: operación de resumen aplicada a datos
- Generalización: jerarquías de conceptos, por ejemplo generalizar isósceles, escaleno y equilátero como triángulo.
- Construcción de atributos: obtención de nuevos atributos a partir de la operación de otros atributos, ejemplo área con altura y ancho

Selección de atributos

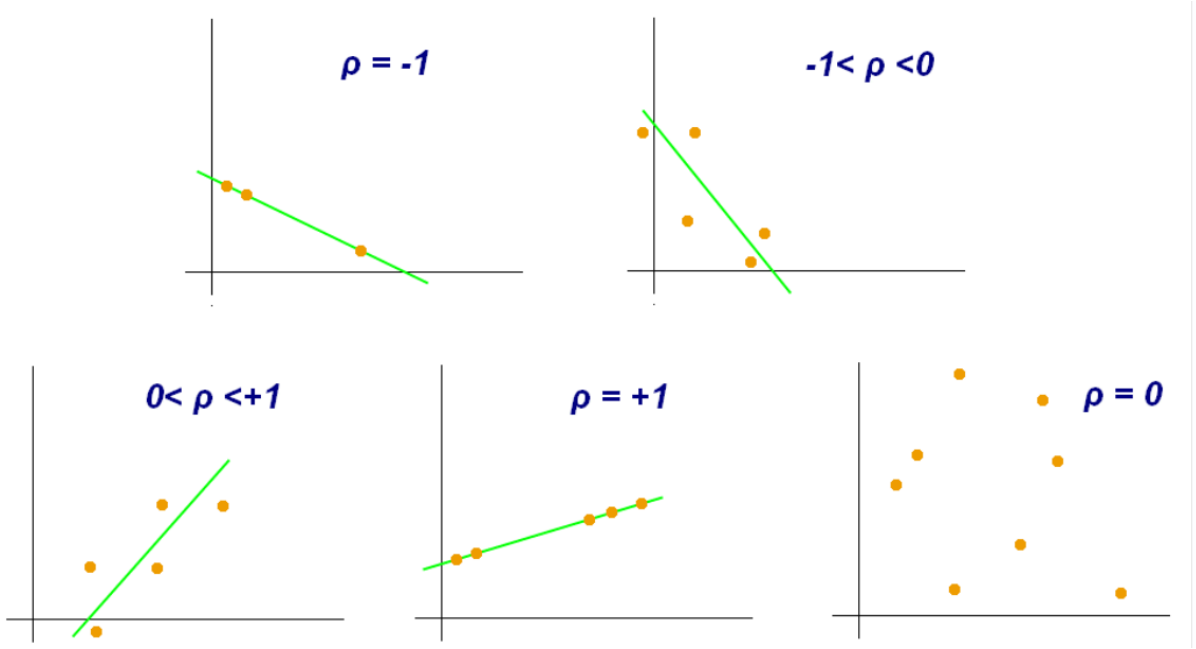
También conocido como selección de características (features) es el proceso de seleccionar los atributos más importantes o relevantes del conjunto de datos. El objetivo es mejorar el rendimiento de predicción mediante la selección de los mejores predictores, es decir aquellos que determinen un mejor rendimiento y calidad del modelo.
Existe diferentes técnicas para la selección de características:

- Métodos de Filtro: la selección de las características se realiza mediante la clasificación según métricas estadísticas que tienden a determinar la correlación de cada característica con la variable de resultado
- Métodos de envoltura: consiste en la selección de un subconjunto y entrenar y evaluar el modelo repetidamente seleccionando en cada ciclo un subconjunto diferente. Básicamente podríamos describirlos como un método de prueba y error. Son muy costosos.
- Métodos integrados: consisten en combinar los métodos de filtro y luego métodos de envoltura.

Siendo los métodos de Filtro los más formales a continuación resumimos en una tabla las métricas a utilizar según el tipo de característica y el tipo de la predicción a realizar:

Predictor / Característica	**Continuo**	**Categórico**
Continuo	Correlación de Pearson	Análisis Discriminante Lineal (LDA)
Categórico	Anova	Chi-Cuadrado

Correlación de Pearson: se usa como una medida para cuantificar la dependencia lineal entre dos variables continuas X e Y, su valor varía de -1 a +1.



Como puede observarse, el signo sólo indica la pendiente de la recta que representa la dependencia lineal. por lo tanto lo que nos importa es el valor absoluto del coeficiente de Pearson, generalmente se utilizan valores de correlación superiores a 0.7:

Valor del coeficiente de correlación	Criterio
De 0,7 a 1,0	Correlación positiva fuerte
De 0,5 a 0,7	Correlación positiva moderada
De 0,2 a 0,5	Correlación positiva baja
De -0,2 a 0,2	Correlación (positiva o negativa) débil o nula
De -0,2 a -0,5	Correlación negativa moderada
De -0,5 a -0,7	Correlación negativa moderada
De -0,7 a -1,0	Correlación negativa fuerte

LDA: el análisis discriminante lineal se usa para encontrar una combinación lineal de características que caracteriza o separa dos o más clases, o niveles, de una variable categórica. Es una generalización del discriminante lineal de Fisher , que busca encontrar una combinación lineal de características que separe en dos o más clases.
Es decir, el objetivo de un LDA suele ser proyectar un espacio de características (un conjunto de datos de muestras n-dimensionales en un subespacio más pequeño k (donde $k \leq n-1$), manteniendo la información de discriminación de clases. En general, la reducción de la dimensionalidad no sólo ayuda a reducir los costes computacionales para una tarea de clasificación determinada, sino que también puede ser útil para evitar el sobreajuste minimizando el error en la estimación de los parámetros.
ANOVA: significa análisis de la varianza y es similar a LDA, excepto por el hecho de que opera mediante una o más funciones independientes categóricas y una función dependiente continua. Proporciona una prueba estadística de si las medias de varios grupos son iguales o no.

Chi-cuadrado: es una prueba estadística que se aplica a los grupos de características categóricas para evaluar la probabilidad de correlación o asociación entre ellos utilizando su distribución de frecuencia. Es decir es una medida de la independencia del campo objetivo respecto al campo de división. Un chi-cuadrado 0 indica concordancia exacta entre ambas los frecuencias observadas y las frecuencias esperadas. Por lo tanto indicaría que las variables son independientes. Por el contrario, un chi-cuadrado elevado normalmente está relacionado con una asociación de las variables.

Al reducir la cantidad de atributos mediante los métodos de selección se produce lo que se conoce como reducción horizontal del conjunto de datos. Es una forma de reducir la dimensionalidad de los datos.

Reducción de datos

Las tareas de limpieza habitualmente ocasionan una reducción de los datos al eliminar instancias inconsistentes o con valores outliers. Se trata de una reducción de instancias, es decir registros o filas, también conocido como reducción vertical.

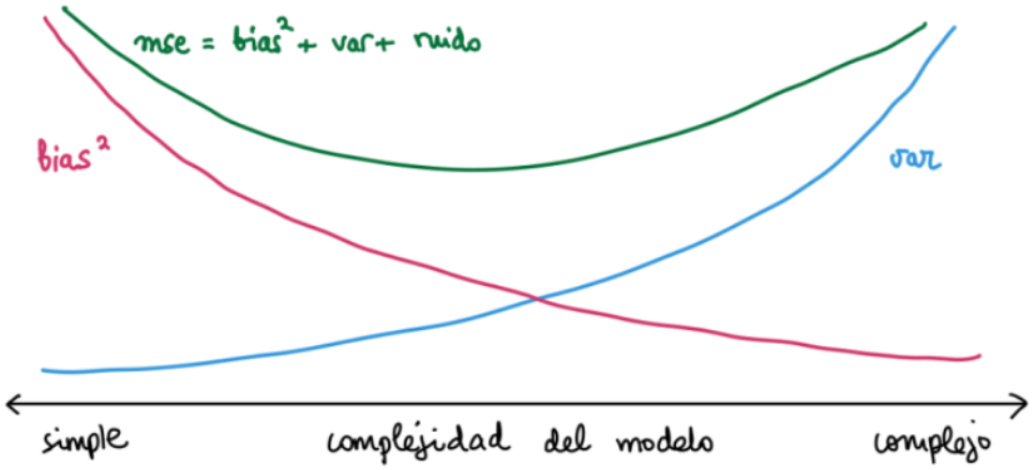
Entrenamiento, Evaluación y Mejora

El entrenamiento consiste en la ejecución del modelo para que aprenda de los datos.
Es importante destacar que las tareas de entrenamiento difieren según el tipo de aprendizaje automático, es decir es diferente para un algoritmo supervisado que para un algoritmo no supervisado o por refuerzo.
No puede verse totalmente independiente de la evaluación y la mejora del modelo. El modelo se entrena, se evalúa mediante diferentes métricas y, en caso de que los resultados de la evaluación no sean aceptables se mejora el modelo volviendo hasta cualquiera de las etapas del proceso (incluída la definición del problema).
Las diferentes formas de particionamiento, los problemas que pueden presentarse en el entrenamiento como tambièn las diferentes métricas de evaluación de modelos serán vistos de manera detallada y aplicadas en las subsiguientes unidades.

Sesgo vs Varianza

Entrenar un modelo de machine learning es abstraer los datos que tenemos para generar predicciones sobre datos que el modelo no ha visto previamente. Esto es encontrar la distribución que se oculta tras los datos y usarla para generar predicciones. El error del modelo es la diferencia entre estos dos valores y entender cómo minimizarlo nos ayuda a construir modelos más robustos y no caer en los problemas clásicos de overfitting o underfitting.
Este error se descompone en sesgo, varianza y ruido. El componente de error debido al ruido es irreducible, por lo tanto lo que se intentará reducir es el sesgo y la varianza.
El objetivo es un modelo que tenga un buen desempeño frente a datos nuevos, capaz de generar predicciones basadas en una correcta generalización a partir de los datos de entrenamiento.
Para medir este desempeño en la práctica, se divide el conjunto de datos disponible entre datos destinados al entrenamiento y datos de test. De esta manera, podremos calcular el

error de manera justa evaluando el modelo con un conjunto de datos que no ha visto durante el entrenamiento.
El sesgo (o bias) es la diferencia entre el valor medio predicho por el modelo y el valor medio real. Si la diferencia entre estas dos magnitudes es elevada, estamos ante un modelo demasiado simple que no ha aprendido las relaciones relevantes entre las variables disponibles y la variable a predecir. A este fenómeno se le llama underfitting o subajuste y los errores de entrenamiento y test serán altos. Un ejemplo de modelo demasiado simple sería ajustar un comportamiento complejo y no lineal a una línea recta



Bibliografía y fuentes de información

https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf

Fuentes:
https://www.ibm.com/mx-es/topics/supervised-learning

https://www.ibm.com/es-es/topics/unsupervised-learning