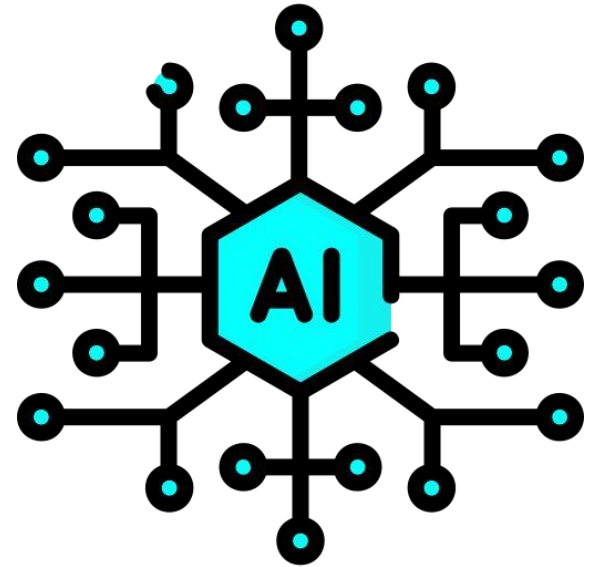


Modelos Predictivos y Machine Learning



Machine Learning

Objetivo 1

¿Qué es la IA?



“Es el campo de la ciencia de computación dedicado a la **resolución de problemas cognitivos asociados comúnmente a la inteligencia humana**, como el aprendizaje, la creación y el reconocimiento de imágenes.”



“Procesos informáticos y los algoritmos **estadísticos que pueden simular y potenciar la inteligencia humana.**”

“Describe sistemas capaces de adquirir conocimientos y aplicarlos a la resolución de problemas.”



“Es un **campo de la ciencia relacionado con la creación de computadoras y máquinas que pueden razonar**, aprender y actuar de una manera que normalmente **requeriría inteligencia humana o que involucre datos cuya escala exceda lo que los humanos pueden analizar.**”

¿Qué es ML?



Inteligencia artificial (IA)

Cualquier técnica que permita a las computadoras imitar el comportamiento humano.



Chatbots



Sistemas basados en reglas



Robots

Aprendizaje automático (ML)

Técnicas de IA que dan a las computadoras la capacidad de aprender sin estar explícitamente programadas para hacerlo.



Decision trees



K-vecinos más cercanos



Regresión Logística

Aprendizaje Profundo (DL)

Un subconjunto de ML, que hace factible el cálculo de redes neuronales multicapa (simulan el comportamiento cerebro humano).

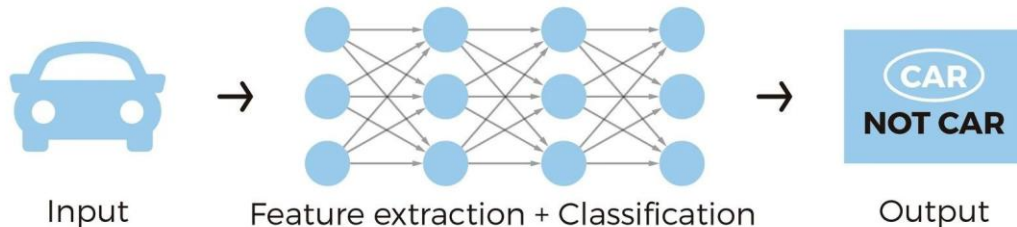


ML VS DL

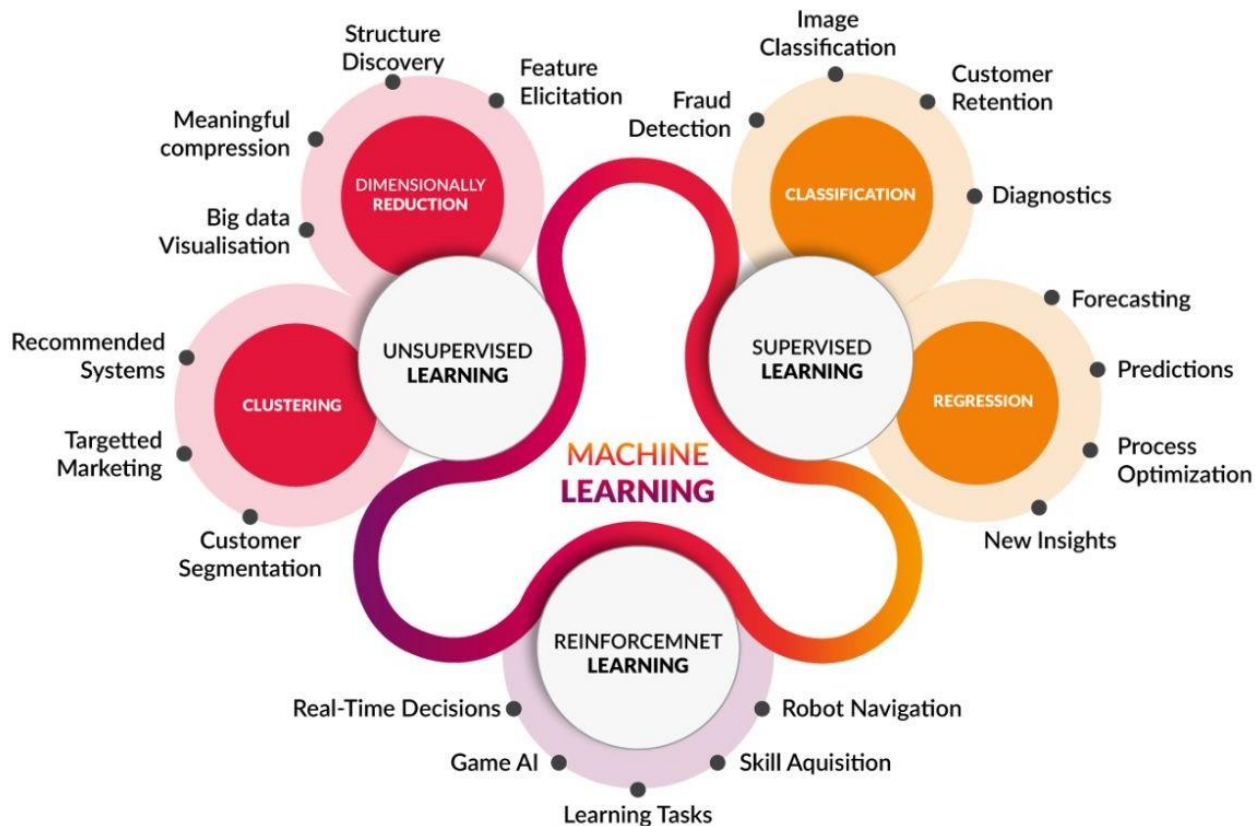
Machine Learning



Deep Learning



Tipos de ML

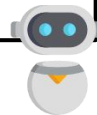


Herramientas para ML

Herramientas para ML

Existe una **infinidad de herramientas a usar y se pueden segmentar por etapa de proyecto de ML:** extracción de la data, limpieza de la data, modelamiento y despliegue. No obstante, actualmente, la apuesta está en el uso de las nubes y, dentro de estas, existe dos posibilidades:

Uso de modelos customizados: implica hacer un desarrollo desde cero.







Modelos pre entrenados: basta con subir nuestra data para obtener resultados.



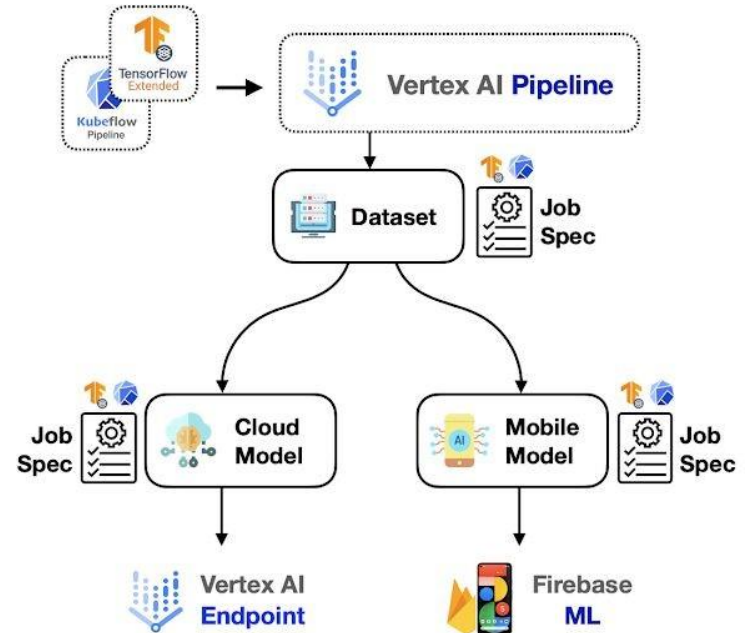
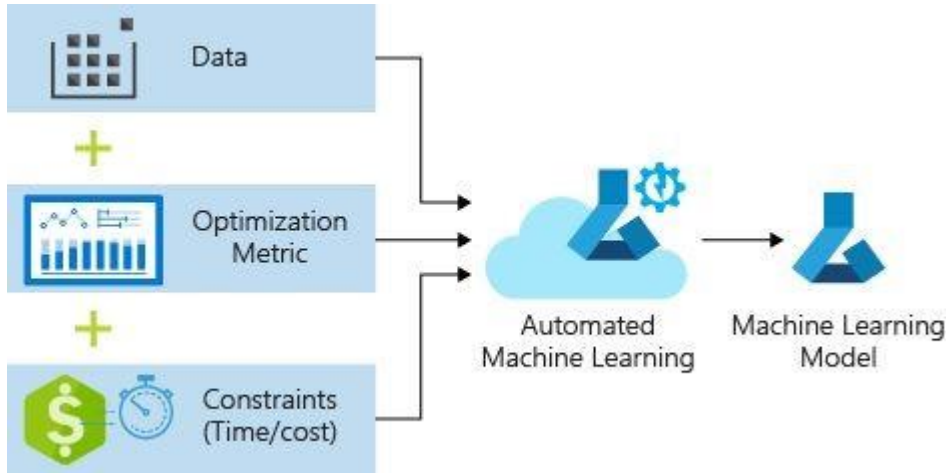
Herramientas para ML

Modelos customizados

				
Machine learning	Amazon SageMaker	Azure Machine Learning	Google Cloud AI Platform	
Image recognition	Amazon Rekognition	Azure Cognitive Services	Google Cloud Vision	
Speech	Amazon Polly, Amazon Transcribe	Azure Cognitive Services	Google Cloud Speech-to-Text and Text-to-Speech	
Natural language processing	Amazon Comprehend	Azure Cognitive Services	Google Cloud Natural Language API:	
Big Data Analytics	Databricks	Databricks	Databricks	
Chat Bot	AWS Chatbot	Azure Bot Service	Dialogflow	
Pricing	Per hour	Per minute	Per minute	

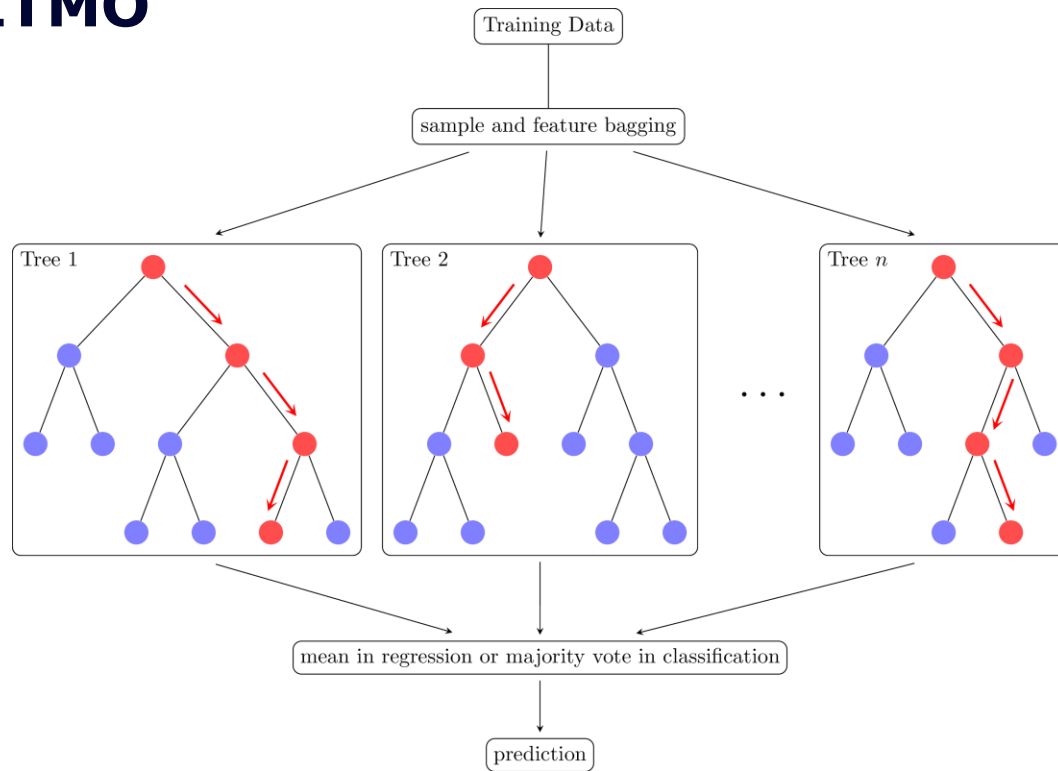
Herramientas para ML

Modelos pre entrenados



RANDOM FOREST

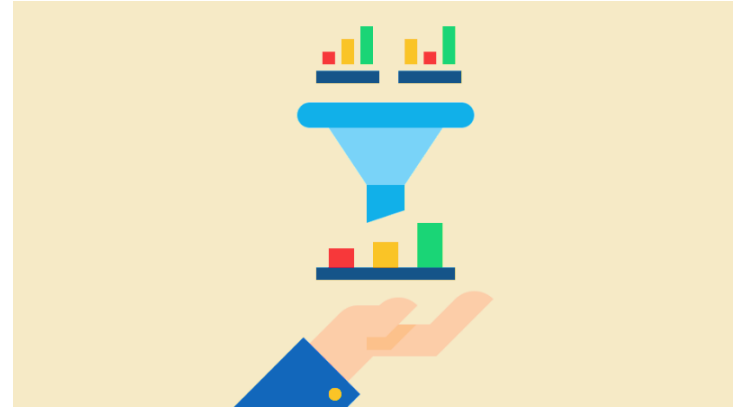
ALGORITMO



PASOS PREVIOS

PREPROCESAMIENTO DE DATOS

1. Detección y eliminación de datos duplicados
2. Manejo de valores nulos (NaN)
3. Conversión de tipos de datos:
4. Codificación de variables categóricas
5. Detección y manejo de *outliers*
6. Normalización o estandarización de datos numéricos (si es necesario)
7. Eliminar o transformar variables irrelevantes
8. Generación de nuevas características (*Feature Engineering*)
9. Balanceo de clases (solo si aplicable)




CODIFICACIÓN DE VARIABLES CATEGÓRICAS

One-Hot Encoding

Crea una nueva columna para cada categoría de la variable. Cada columna tiene un valor binario (0 o 1), donde 1 indica la presencia de la categoría en esa fila y 0 su ausencia.

Es útil cuando **las categorías no tienen un orden inherente**

Colour	
Green	
Red	
Blue	




Green	Red	Blue
0	1	1
1	1	1
1	0	1
0	0	0
0	1	0

Label Encoding

Asigna un valor numérico entero único a cada categoría en la variable. Esencialmente, cada categoría de la variable se transforma en un número.

Es ideal cuando **existe un orden inherente** en las categorías, ya que introduce una relación implícita entre los valores numéricos asignados

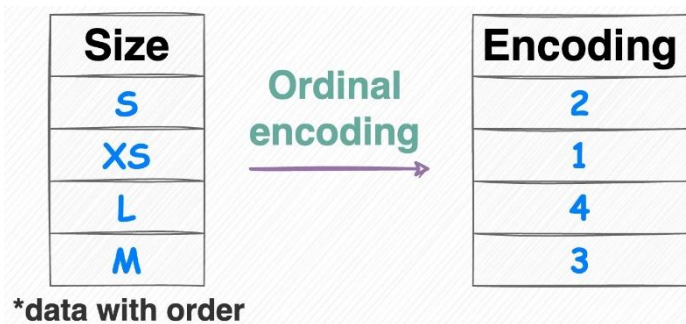
Original Data			Label Encoded Data	
Team	Points		Team	Points
A	25		0	25
A	12		0	12
B	15		1	15
B	14		1	14
B	19		1	19
B	23		1	23
C	25		2	25
C	29		2	29

CODIFICACIÓN DE VARIABLES CATEGÓRICAS

Ordinal Encoding

Similar a Label Encoding, asigna un número a cada categoría. Sin embargo, a diferencia de Label Encoding, se utiliza cuando hay un **orden específico y definido** entre las categorías, que el modelo debe reconocer.

Se aplica cuando se trabaja con **variables categóricas ordinales** y el orden entre las categorías tiene un significado importante



pd.get_dummies()



LabelEncoder()

OneHotEncoder()

OrdinalEncoder()

SELECCIÓN DE FEATURES

La selección de características (*feature selection*) tiene el fin de identificar las características que tienen mayor relevancia en la predicción.

Se puede realizar la selección en base a distintos criterios, sin embargo, los más utilizados y destacables pueden ser:

- ☐ Análisis Exploratorio Inicial
- ☐ Matriz de Correlación
- ☐ Selección Basada en Importancia Estadística
- ☐ Selección Basada en Algoritmos de Aprendizaje Automático
- ☐ Selección Basada en Reducción de Dimensionalidad

Es muy común separar la variable objetivo con las características a analizar

- La variable objetivo (**target**) será la columna que contiene el precio de las casas.
- Las características (**features**) serán todas las demás columnas que consideres relevantes para predecir el precio.

MATRIZ DE CORRELACIÓN

Esta matriz es una tabla que muestra los coeficientes de correlación entre las variables.

Es una herramienta clave en el análisis de datos, ya que permite identificar la relación entre las diferentes características del dataset

COEFICIENTE DE CORRELACIÓN DE PEARSON (Valor entre -1y 1)

Mide la intensidad y dirección de una relación lineal entre dos variables y es útil para detectar **multicolinealidad**, que ocurre cuando dos o más características están altamente correlacionadas entre sí, lo cual puede afectar el rendimiento de ciertos modelos, como las regresiones lineales

- Valores positivos indican una correlación directa, es decir, cuando una variable aumenta, la otra también lo hace.
- Valores negativos indican una correlación inversa, es decir, cuando una variable aumenta, la otra disminuye.
- Valores cercanos a 0 indican que no hay una correlación lineal significativa.

DIVISIÓN DEL DATASET

Dividir el dataset en **conjuntos de entrenamiento y prueba** es fundamental para evaluar el rendimiento del modelo. Se entrena el modelo con los datos de entrenamiento y luego se prueba en datos que no ha visto antes (conjunto de prueba) para determinar si generaliza bien

La proporción más común es dividir el dataset en un **80% para entrenamiento y un 20% para prueba**. También se podría usar un conjunto de validación adicional para ajustar hiperparámetros, aunque para la mayoría de los casos train_test_split es suficiente

X: CARACTERÍSTICAS (*FEATURES*)

Y: VARIABLE OBJETIVO (*TARGET*)

