Ciencia de Datos

- **Módulo 2**

Variable categóricas y numércias

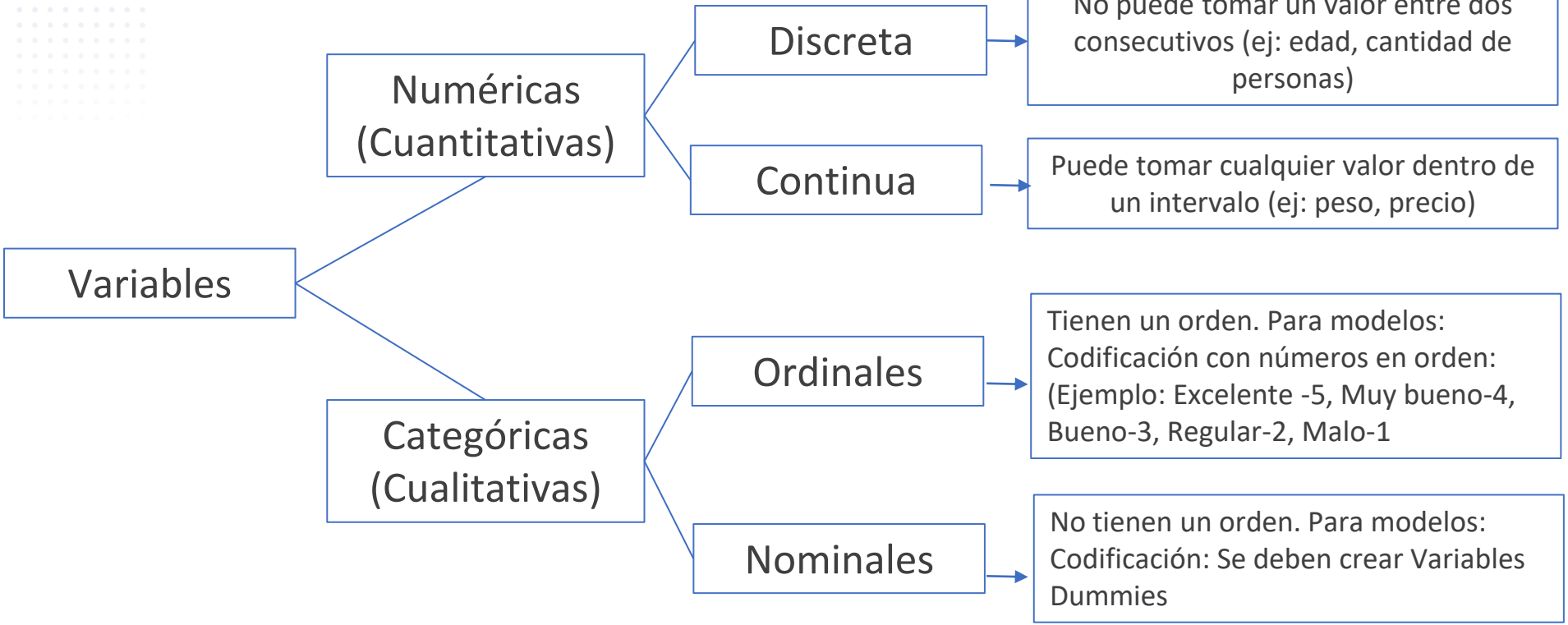
UTNFra

GyEE Gestión y Evaluación Educativa

<aprendé programando Virtual>

SECUNDARIA del futuro

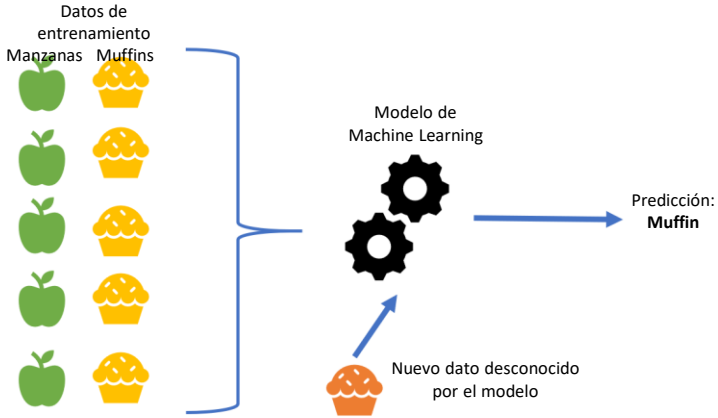
Variables


Variables
Numéricas (Cuantitativas)
Discreta
No puede tomar un valor entre dos consecutivos (ej: edad, cantidad de personas)
Continua
Puede tomar cualquier valor dentro de un intervalo (ej: peso, precio)
Categóricas (Cualitativas)
Ordinales
Tienen un orden. Para modelos: Codificación con números en orden: (Ejemplo: Excelente -5, Muy bueno-4, Bueno-3, Regular-2, Malo-1
Nominales
No tienen un orden. Para modelos: Codificación: Se deben crear Variables Dummies

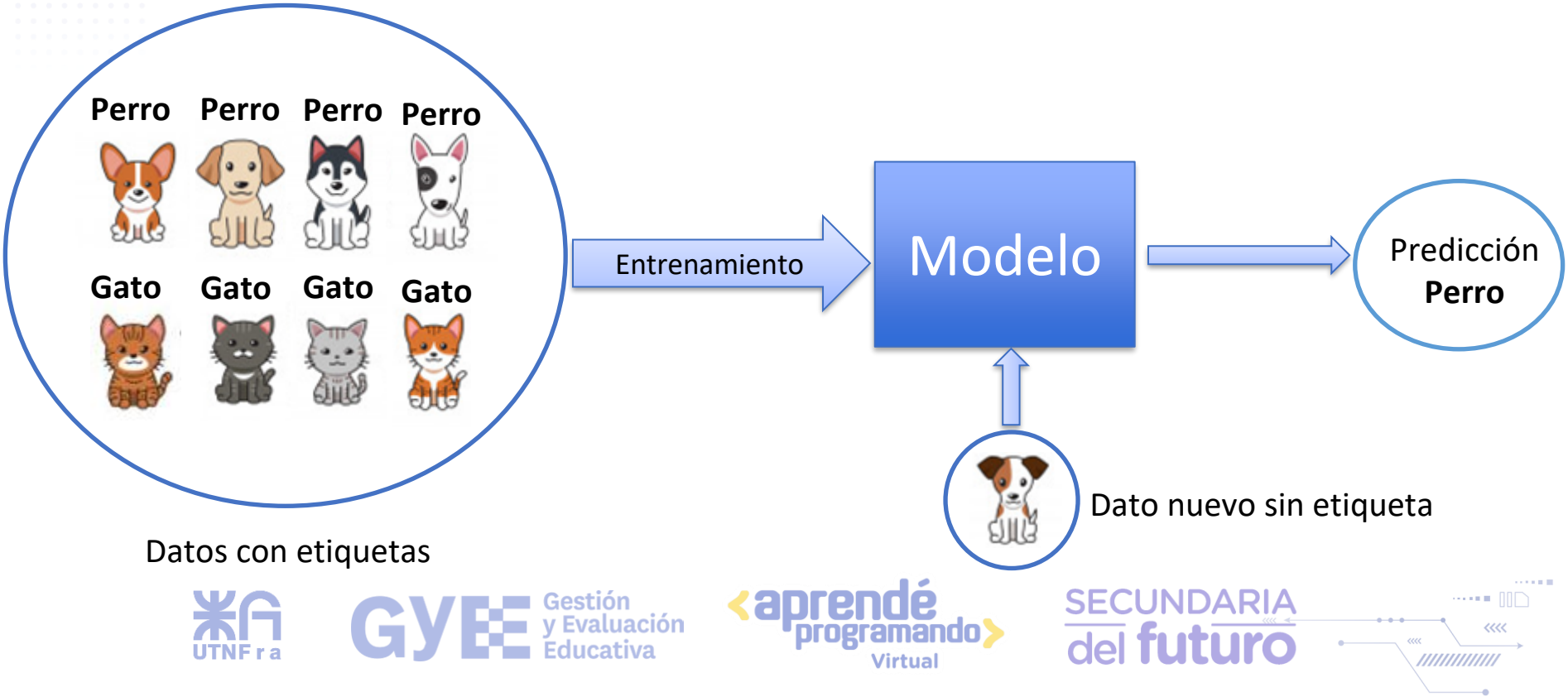
Aprendizaje Automático - Tipos

Supervisado: Se dice que un modelo de ML es supervisado cuando, los datos con los que se entrena el modelo tienen los valores que queremos predecir, a estos se le dice que están etiquetados.

Por ejemplo, si queremos predecir cuántos goles marcará nuestro equipo en base a datos de partidos anteriores, estos datos deben tener cuántos goles marcó nuestro equipo en cada uno de ellos.



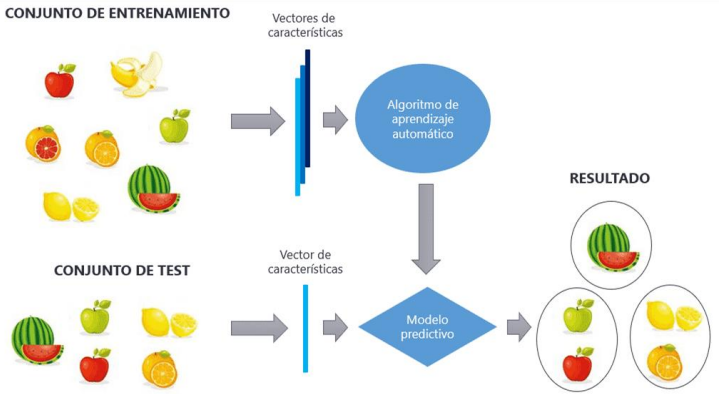
Aprendizaje supervisado



Aprendizaje Automático - Tipos

No supervisado: Estos modelos se entrenan sin información sobre el atributo que se quiere predecir, por lo que la evaluación de su desempeño es más compleja. Sueles usarse para realizar agrupaciones o clusters y como métodos de reducción de dimensionalidad.

Por ejemplo, cuando una compañía quiere hacer grupos de clientes según sus gastos mensuales, cantidades de compras, visitas a la tienda, etc.

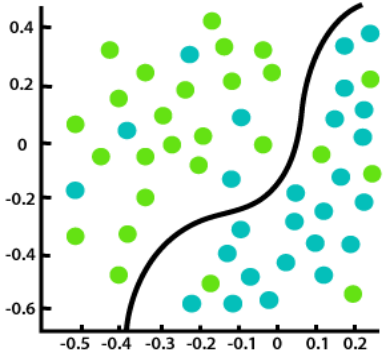


Fuente: https://www.diegocalvo.es/aprendizaje-no-supervisado/

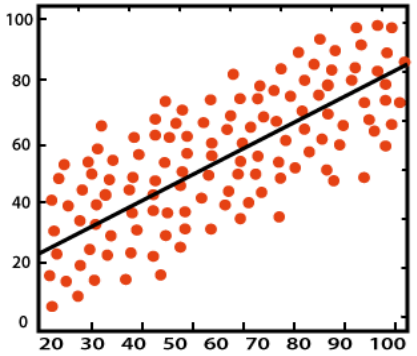
Aprendizaje Automático Supervisado - Tipos

Los principales usos de los modelos supervisados son:

- **Clasificación**: Cuando la variable a predecir es una clase (variable categórica), por ejemplo: Enfermo o no enfermo, la raza de la foto de un animal, etc.
- **Regresión**: Cuando la variable a predecir es un valor (numérica), por ejemplo: Precio de un objeto, nota de un estudiante, probabilidad de lluvia, etc.



Clasificación

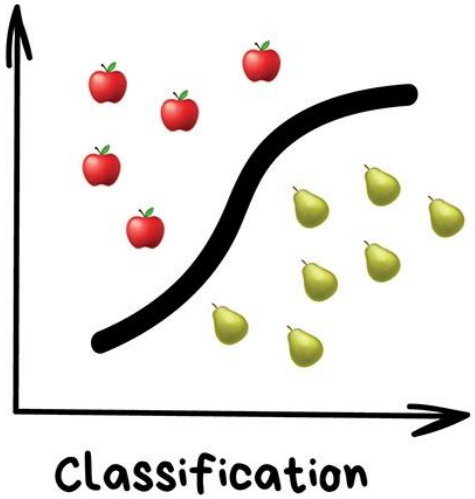


Regresión

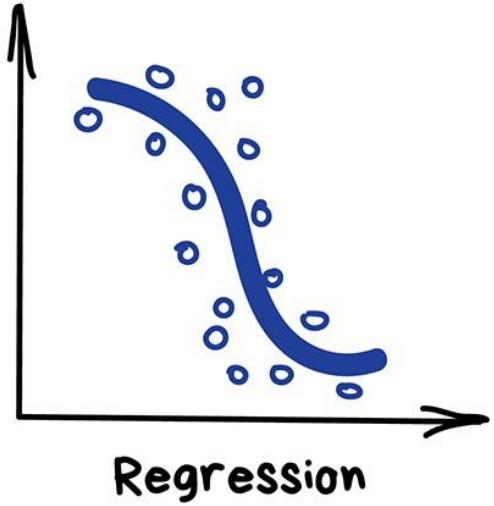
Fuente: https://www.javatpoint.com/regression-vs-classification-in-machine-learning

Clasificación - Regresión

Etiqueta cualitativa



Etiqueta cuantitativa

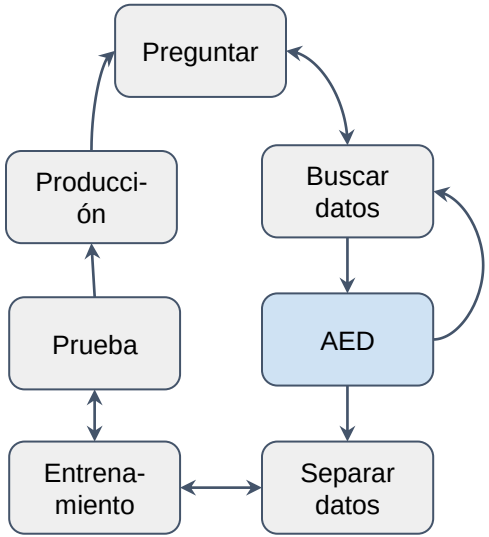


Fuente:https://www.juanbarrios.com/inteligencia-artificial-y-machine-learning-para-todos

Aprendizaje Automático - Pasos

En todo modelo de Machine Learning hay algunos **pasos comunes** que nos sirven de guia:

- **Definición del problema**: Siempre que se va a realizar un proyecto de ML es para resolver un problema específico, por ende, primero debemos de realizarnos una pregunta la cual buscaremos resolver.
- **Buscar los datos**: Por lo general debemos definir qué datos necesitamos y de dónde los podemos encontrar. Hoy en día hay muchos datos libres, sean generados por organismos gubernamentales u ONG, o en algunos casos incluso empresas privadas.
- **Análisis exploratorio de datos**: Durante esta etapa se *limpian* los datos y se realiza una exploración de los mismos, buscando características y relaciones que nos ayuden a seleccionar qué variables usar o que modelo usar.

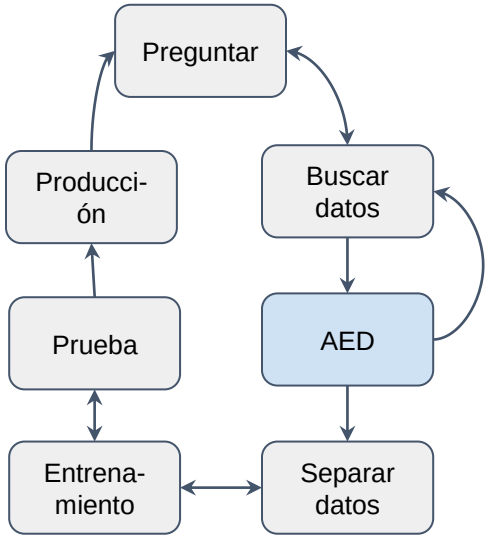


Aprendizaje Automático - Pasos

- **Dividir los datos de entrenamiento y prueba**. Esta división es de vital importancia para poder definir si el modelo tiene un buen rendimiento o no.

- **Entrenamiento del modelo**: En esta etapa se **entrenan** diferentes de modelos y se **prueban** tomando en cuenta alguna métrica previamente definida, seleccionando el modelo que mejor desempeño tenga.

- **Puesta en producción**: Finalmente se realiza la puesta en producción, este paso depende de qué clase de proyecto sea, puede que en algunos casos, colocar en producción solo signifique generar las predicciones para alguna colección de datos, o puede ser implementarlo en un sistema que reciba datos periódicamente.



Datos para entrenamiento del modelo

Para finalizar la Exploración y Limpieza de datos es necesario tener en cuenta que en el entrenamiento de un modelo se realizan operaciones matemáticas por lo cual no puede haber variables categóricas y deben deben convertirse en variables numéricas:

- Cuando la variable categórica es de tipo **ordinal**, es decir, tiene un orden, podemos reemplazar sus valores con números consecutivos. Por ejemplo: Excelente: 5, Muy bueno: 4, Bueno: 3, Regular: 2 y Malo:1.
- Cuando la variable categórica es de tipo **nominal**, es decir, no tiene un orden debemos convertirlas en **variables dummie**

Variables dummies

Variables dummies: Variable ficticias que solo puede tomar como valor 0 y 1. Se utiliza para indicar la presencia o ausencia de valores categóricos

Al tener una variable categórica nominal: se generan columnas de tal manera que cada columna es una categoría que tiene como valor 0 y 1

Ejemplo. Variable Descripción en el dataset Clima:
En la Notebook lo realizaremos

Python:
Función de pandas get_*dummies*
o crear una función
(también hay implementación de Scikit-learn: OneHotEncoder)

	descripcion		Cold	Normal	Warm
0	Cold	0	1	0	0
1	Warm	1	0	0	1
2	Normal	2	0	1	0
3	Cold	3	1	0	0
4	Cold	4	1	0	0