Ayudantía 9

ML, KNN y Árboles

Por Bernardita Alliende y Carlos Stappung

18 de octubre 2024

Contenidos

- Terminología Machine Learning
- Entrenamiento de Modelos
- KNN
- Árboles y ensambles
- Ejemplo de código

Terminología de ML

- Conjunto de datos (dataset)

id	Tamaño	Peso	Color	Tipo
1	38	20	Negro	Perro
2	40	14	Café	Perro
3	20	5	Naranjo	Gato
4	23	4	Negro	Gato

Terminología de ML

- Dato/Instancia

id	Tamaño	Peso	Color	Tipo
1	**38**	**20**	**Negro**	**Perro**
2	40	14	Café	Perro
3	20	5	Naranjo	Gato
4	23	4	Negro	Gato

Terminología de ML

- Atributos/Características/Variables/Columnas

id	Tamaño	Peso	Color	Tipo
1	38	20	Negro	Perro
2	40	14	Café	Perro
3	20	5	Naranjo	Gato
4	23	4	Negro	Gato

Terminología de ML

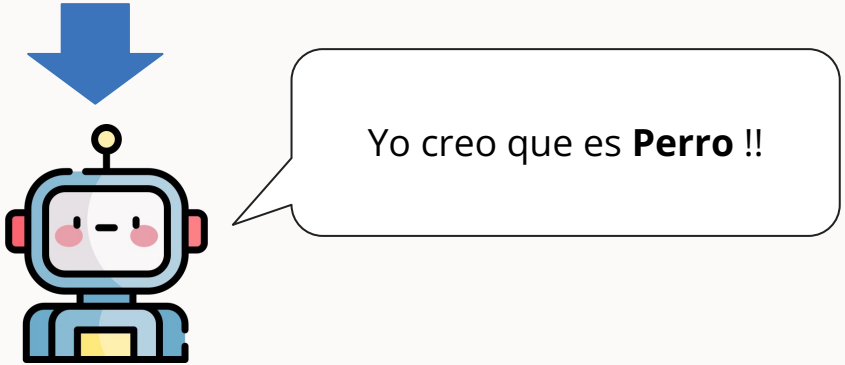
- Objetivo/Target/Salida/Clase

id	Tamaño	Peso	Color	**Tipo**
1	38	20	Negro	**Perro**
2	40	14	Café	**Perro**
3	20	5	Naranjo	**Gato**
4	23	4	Negro	**Gato**

Terminología de ML

- Predicción

id	Tamaño	Peso	Color	Tipo
1	38	20	Negro	**?**



Terminología de ML

- Etiqueta

Etiqueta 1: **Perro**

Etiqueta 2: **Gato**

Terminología Machine Learning

- Parámetro

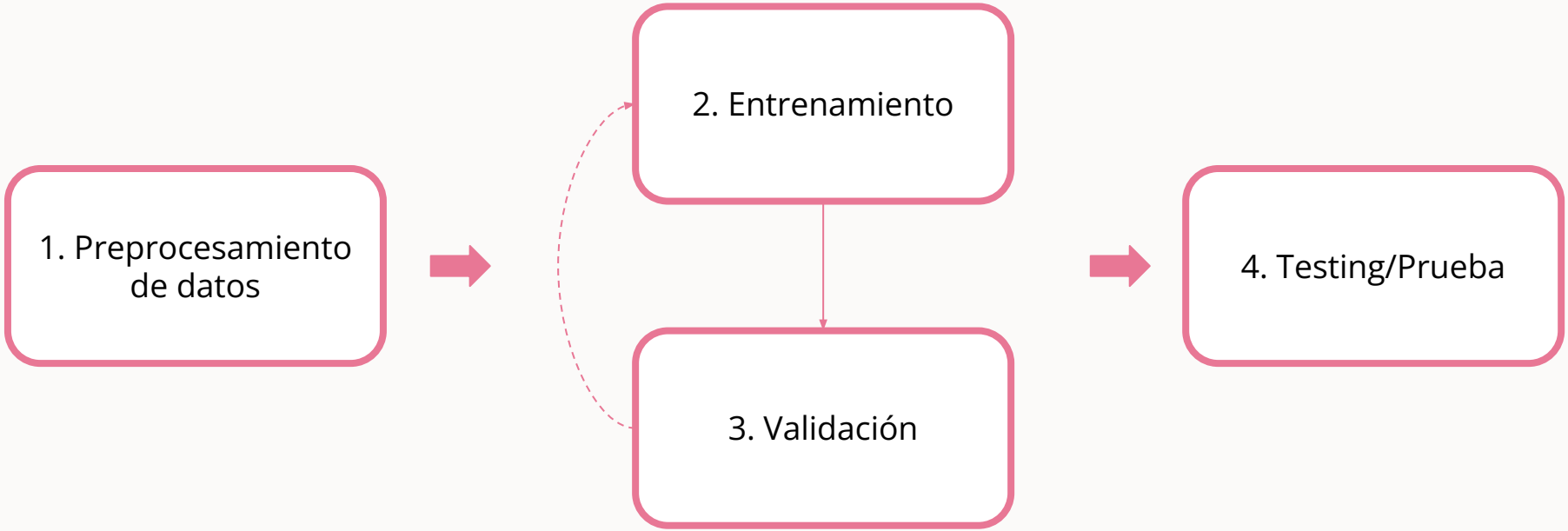
Valores internos que el modelo ajusta (solito, nosotros no hacemos nada) durante el proceso de entrenamiento para minimizar el error del modelo.

Terminología Machine Learning

- Hiperparámetro

Valores que nosotros podemos modificar para que el modelo tenga un mejor rendimiento.

Entrenamiento Modelo



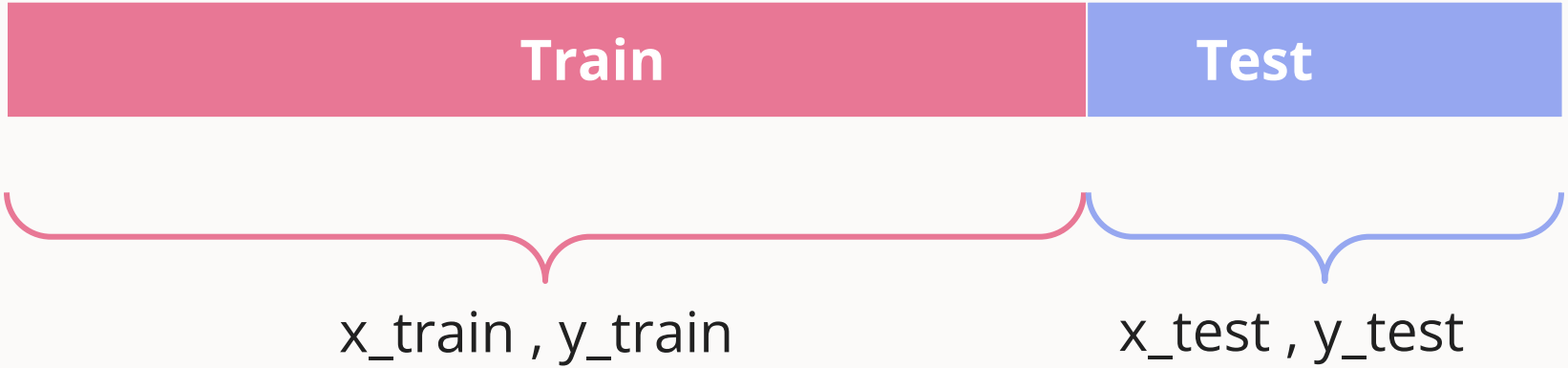
1. Preprocesamiento de datos

- La calidad de los datos impacta en la efectividad del modelo
- Asegurar que los datos estén limpios, completos, y del formato necesario
- Mejorar el rendimiento

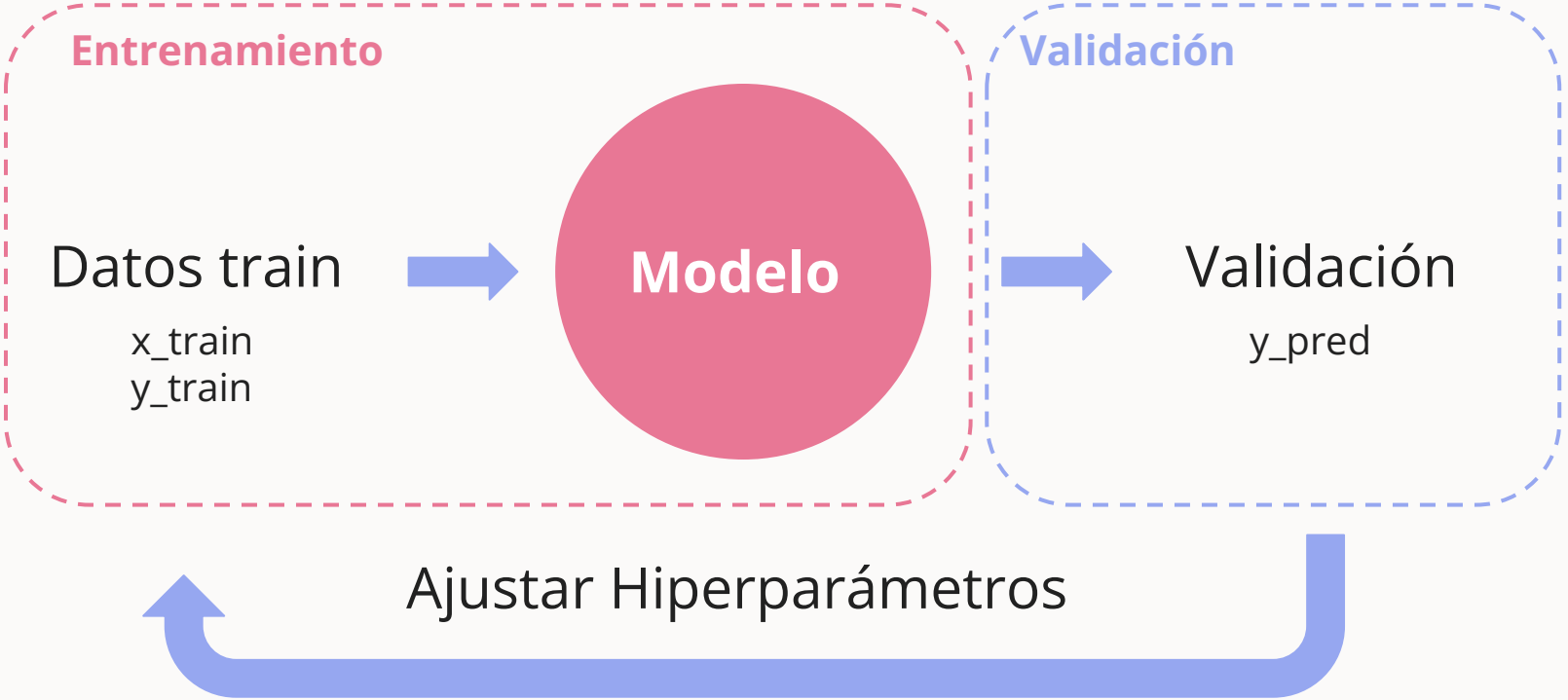
1. Preprocesamiento de datos

- Normalización y Estandarización de los datos
- Codificación de los datos
- Reducción de dimensionalidad, extrayendo características
- Imputación para reemplazar datos faltantes

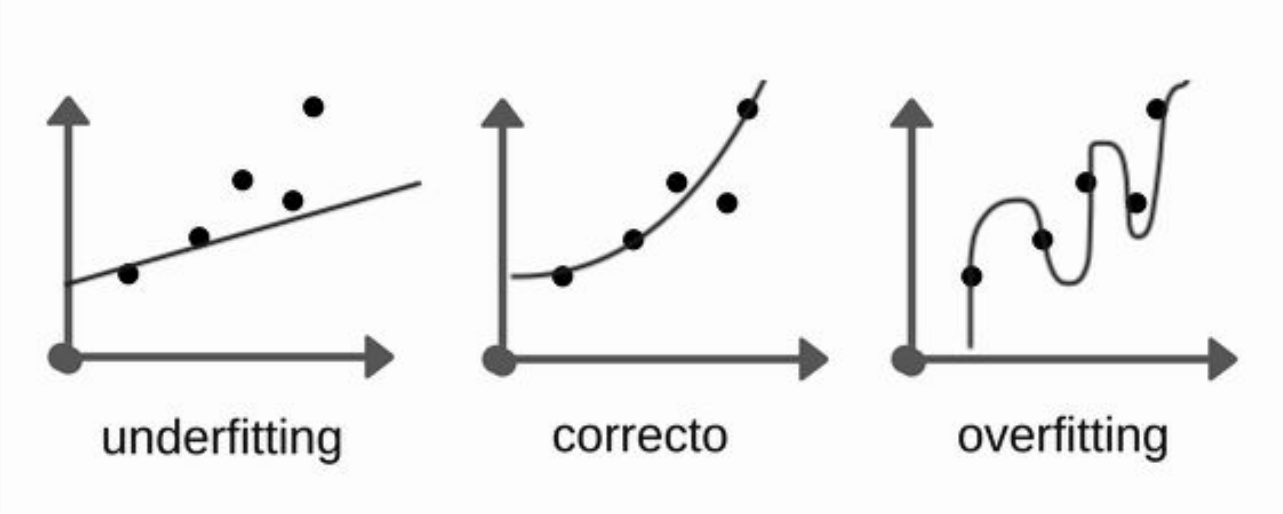
Separación de los Datos



2. Entrenamiento y 3. Validación



Sobreajuste (overfitting)



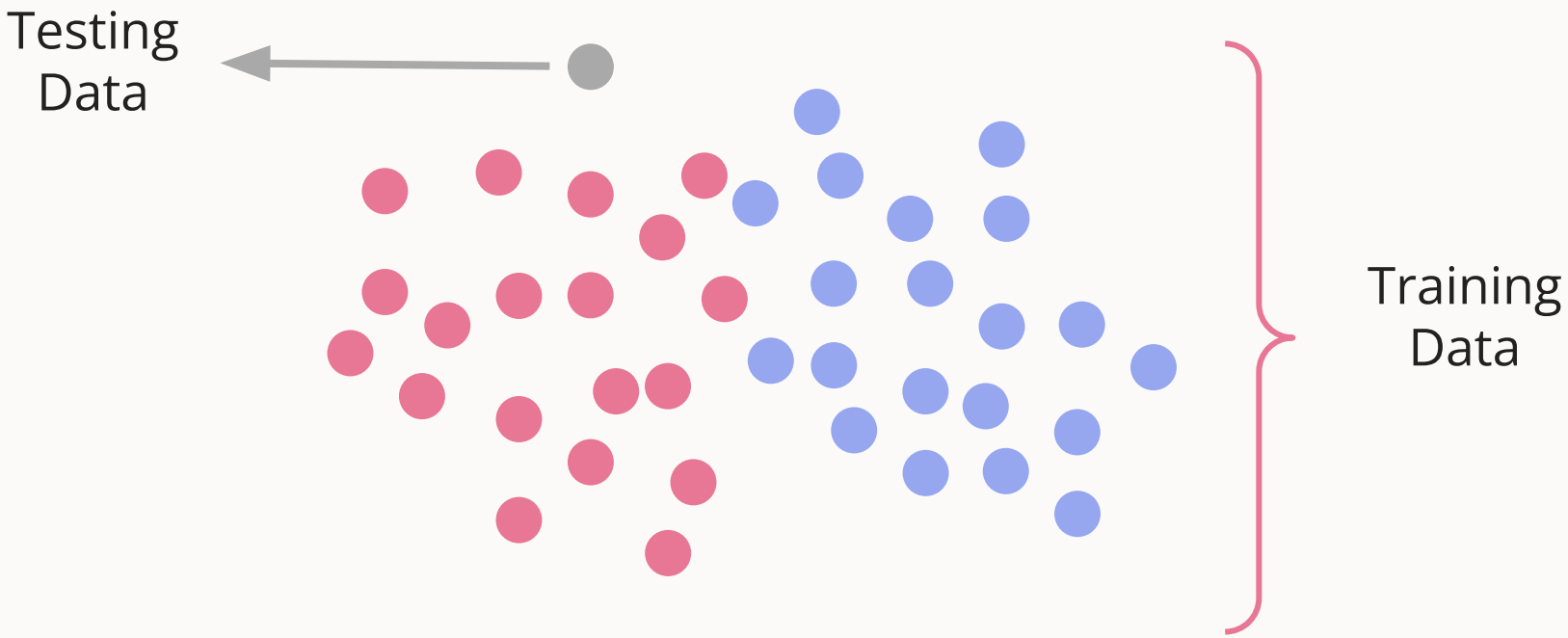
4. Testeo

- Proceso de evaluar la precisión final del modelo usando un conjunto de datos separado que no se usó durante el entrenamiento o la validación

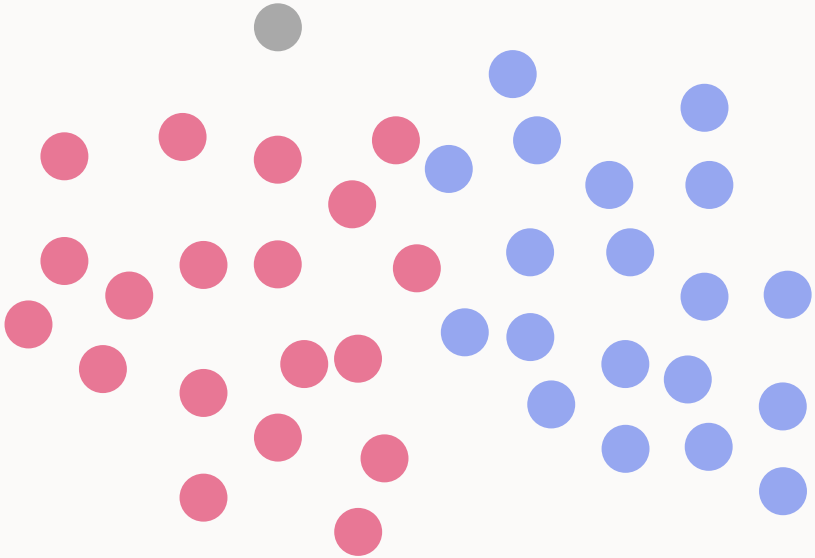
Algoritmo: K-Nearest Neighbor (KNN)

- Clasificación de datos nuevos
- No hay entrenamiento como tal
- De los algoritmos más simples para probar experimentos rápido
- Basado en distancias.

Algoritmo: K-Nearest Neighbor (KNN)



Algoritmo: K-Nearest Neighbor (KNN)

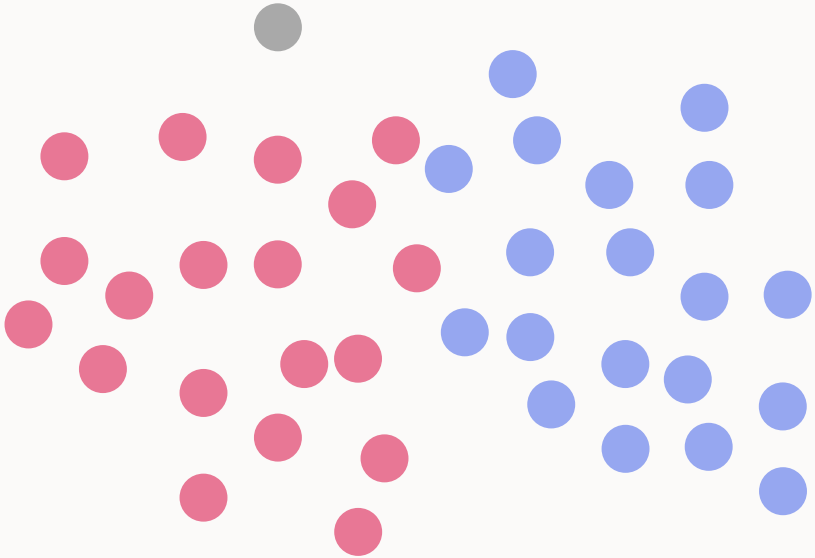


Algoritmo KNN:

1. Distancias desde 

0.7935
0.0278
0.4788
0.0835
0.1108
0.7716
0.6692
0.0075
0.8256
0.8887
0.6573
0.5489
0.3124
0.4698
0.4479

Algoritmo: K-Nearest Neighbor (KNN)

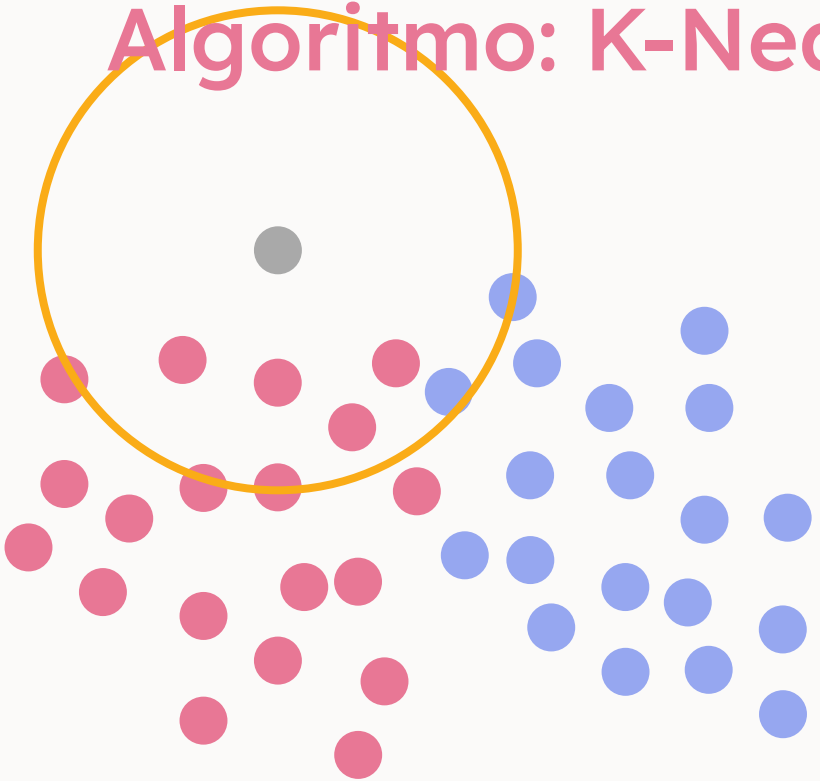


Algoritmo KNN:

1. Distancias desde
2. Ordenarlas

0.0075
0.0278
0.0835
0.1108
0.3124
0.4479
0.4698
0.4788
0.5489
0.6573
0.6692
0.7716
0.7935
0.8256
0.8887

Algoritmo: K-Nearest Neighbor (KNN)



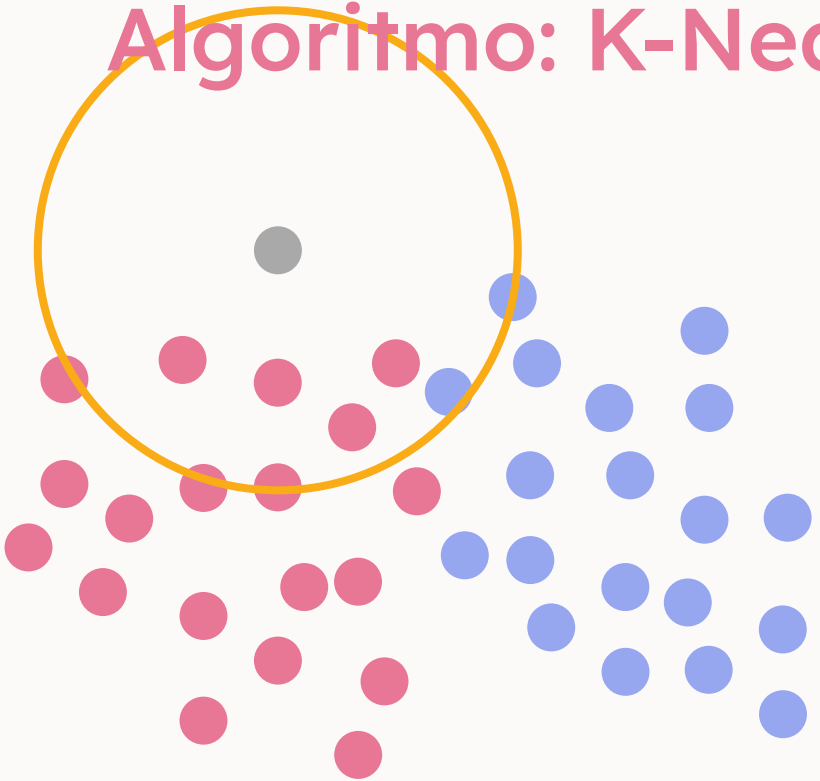
Algoritmo KNN:

1. Distancias desde ●
2. Ordenarlas
3. Seleccionar las K más cercanas

K = 7

0.0075
0.0278
0.0835
0.1108
0.3124
0.4479
0.4698
0.4788
0.5489
0.6573
0.6692
0.7716
0.7935
0.8256
0.8887

Algoritmo: K-Nearest Neighbor (KNN)



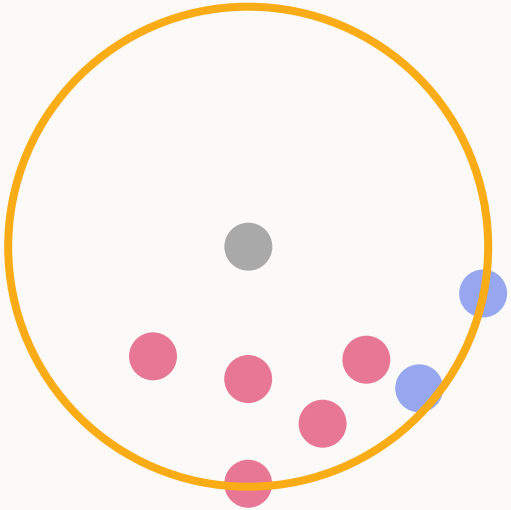
Algoritmo KNN:

1. Distancias desde ●
2. Ordenarlas
3. Seleccionar las K más cercanas
4. Clasificar por mayoría de votos

K = 7

0.0075
0.0278
0.0835
0.1108
0.3124
0.4479
0.4698
0.4788
0.5489
0.6573
0.6692
0.7716
0.7935
0.8256
0.8887

Algoritmo: K-Nearest Neighbor (KNN)



4. Clasificar por mayoría de votos

5 ●

2 ●

Testing Sample is ●

Algoritmo: K-Nearest Neighbor (KNN)

PARÁMETROS

- Conjunto de datos del entrenamiento

HIPERPARÁMETROS

- Número k de vecinos
- Métrica para medir las distancias
- Pesos de los vecinos

Árboles de decisión

- Usados para clasificación y regresión.

- Organizan datos para dividirlos en diferentes clases.

Árboles de decisión

Clima	Temperatura	Humedad	Viento	Jugar?
soleado	alta	alta	F	No
soleado	alta	alta	V	No
nublado	alta	alta	F	Si
lluvioso	Agradable	alta	F	Si
lluvioso	frio	normal	F	Si
lluvioso	frio	normal	V	No
nublado	frio	normal	V	Si
soleado	Agradable	alta	F	No
soleado	frio	normal	F	Si
lluvioso	Agradable	normal	F	Si
soleado	Agradable	normal	V	Si
nublado	Agradable	alta	V	Si
nublado	alta	normal	F	Si
lluvioso	Agradable	alta	V	No

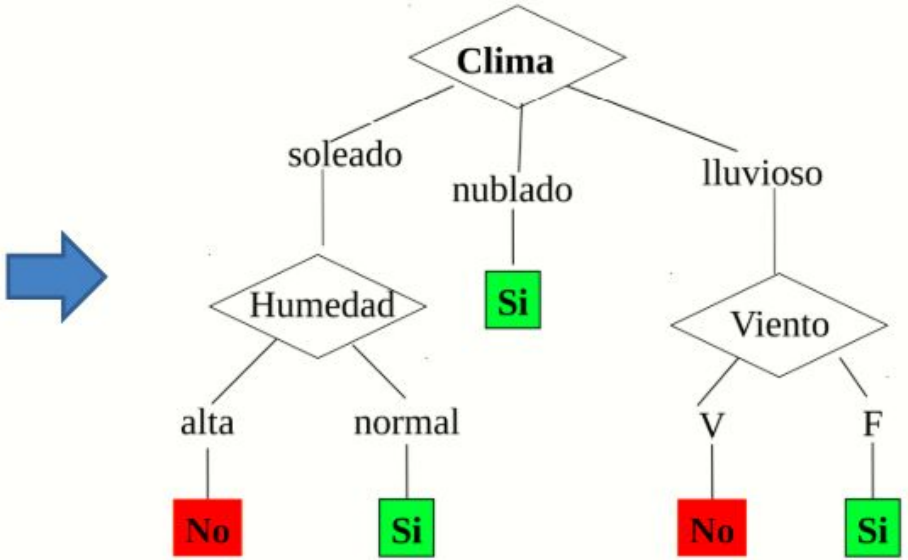


Imagen obtenida de la clase de árboles del profesor Hans Löbel 2023-1

Árboles de decisión

Estructura del Árbol:

- Raíz: Representa la totalidad de los datos.

- Nodos:
  - Cada nodo es una pregunta o condición sobre alguna característica.
  - Según la respuesta, los datos se dividen en nodos del siguiente nivel.

Árboles de decisión

Estructura del Árbol:

- Para decidir la mejor división en un nodo se utilizan métricas como Gini, Entropía o índice de impureza.
- Estas son calculadas por los modelos automáticamente

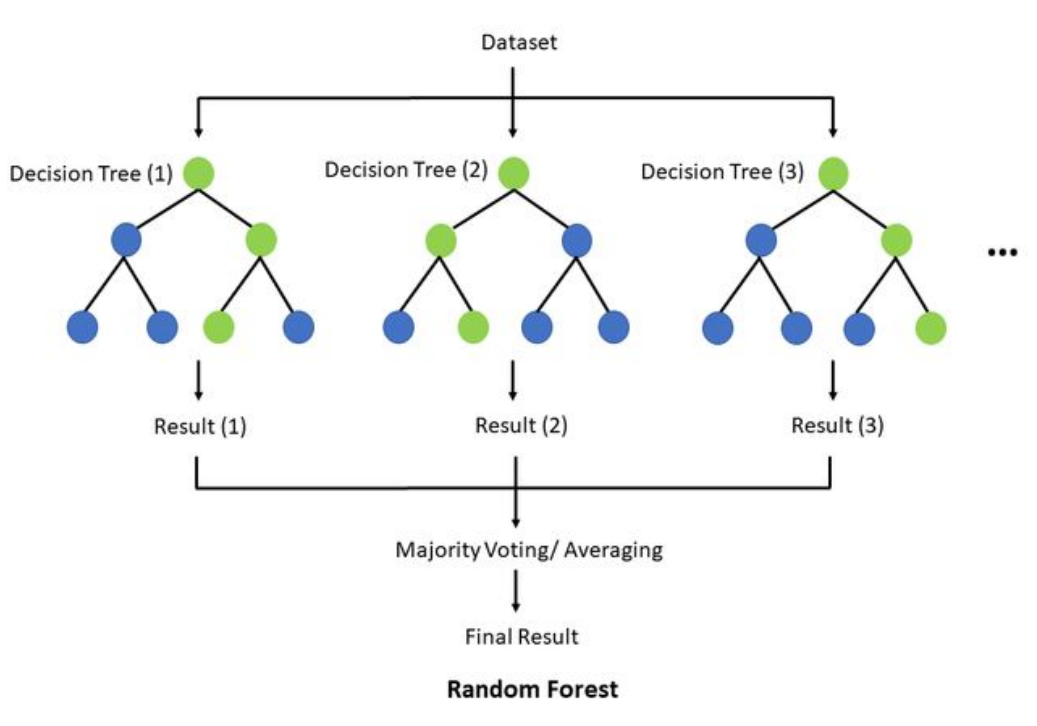
Hiperparámetros: Árboles

- **criterion**: función para calcular la calidad de la división de un nodo (ej. gini, entropy).
- **max_depth**: profundidad máxima del árbol
- **min_samples_split**: número mínimo de muestras requeridas para dividir un nodo
- **min_samples_leaf**: Número mínimo de muestras requeridas en una hoja.

Ensamblajes

- **Random Forest**
- **XGBoosting**

Random Forest



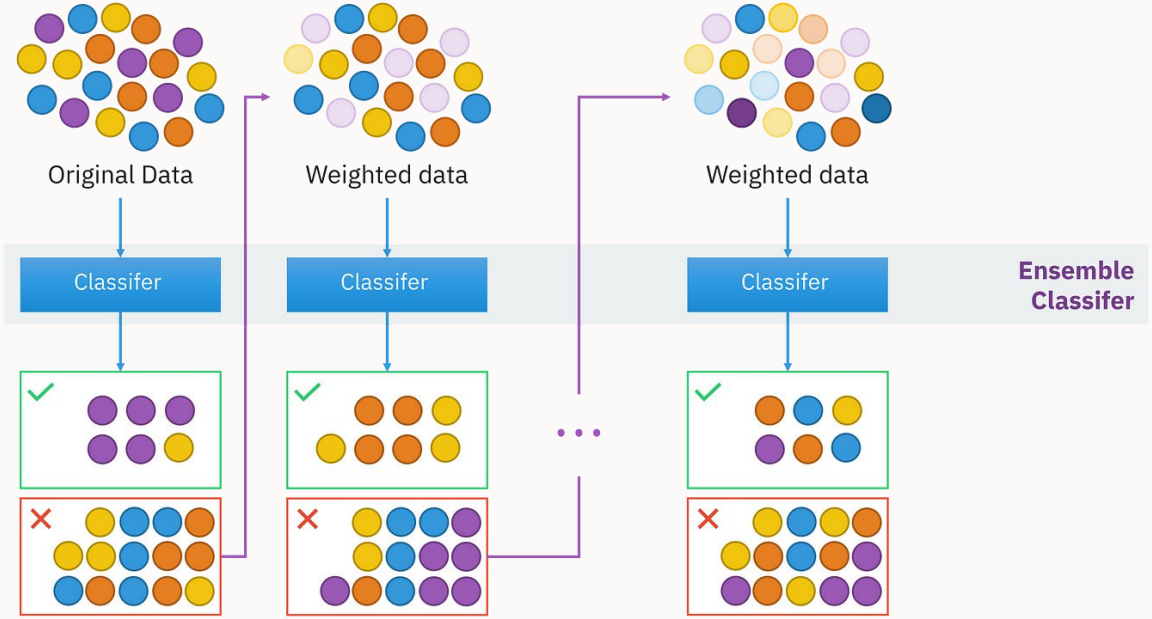
Random Forest

- Un árbol para cada muestra aleatoria (subconjunto) de los datos.
- Árboles de poca profundidad para evitar el overfitting.
- La predicción final se obtiene de un promedio entre las predicciones de los árboles del bosque

Hiperparámetros: RF

- **Los mismos de un árbol**
- **n_estimators:** número de árboles en el bosque

XGBoosting


Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer

XGBoosting

- Árboles secuenciales donde cada árbol corrige los errores de los anteriores

- La predicción final se obtiene por el último árbol

Hiperparámetros: XGB

- **booster:** tipo de modelo base (se puede construir con árboles o con modelos lineales)
- **objective:** función de pérdida que se optimiza (Estima que tanto error tiene el árbol)
- **subsample:** Proporción de muestras a utilizar para entrenar cada modelo

Ayudantía 9

ML, KNN y Árboles

Por Bernardita Alliende y Carlos Stappung

18 de octubre 2024