Ayudantía 8

KNN y Árboles

Por Ignacio Villanueva y Kaina Galdames

20 de mayo 2024

Contenidos

- Terminología Machine Learning
- Preprocesamiento de datos
- KNN
- Árboles y ensambles
- Ejemplo de código

Terminología Machine Learning

- Dato

Una pieza individual de información. Puede ser un número, texto, imagen, etc.

id	Altura	Ancho	Color	Clase
1	12	3	Rojo	Auto
2	34	5	Rojo	Auto
3	16	64	Azul	Moto
4	7	1	Verde	Moto

Terminología Machine Learning

- Conjunto de datos

Colección de datos estructurados, generalmente organizados en filas (instancias) y columnas (atributos/características).

id	Altura	Ancho	Color	Clase
1	12	3	Rojo	Auto
2	34	5	Rojo	Auto
3	16	64	Azul	Moto
4	7	1	Verde	Moto

Terminología Machine Learning

- Atributos/Features/Características/Columnas

Una propiedad o dimensión que describe algún aspecto de los datos. Por ejemplo, en un conjunto de datos de viviendas, podrían ser "tamaño", "número de habitaciones", etc.

id	Altura	Ancho	Color	Clase
1	12	3	Rojo	Auto
2	34	5	Rojo	Auto
3	16	64	Azul	Moto
4	7	1	Verde	Moto

Terminología Machine Learning

- Objetivo

El valor real que queremos predecir con nuestro modelo. También se le puede llamar "target", “clase” o "variable de salida".

id	Altura	Ancho	Color	Clase
1	12	3	Rojo	Auto
2	34	5	Rojo	Auto
3	16	64	Azul	Moto
4	7	1	Verde	Moto

Terminología Machine Learning

- Predicción

El valor que el modelo intenta estimar o predecir.

id	Altura	Ancho	Color	Clase
1	12	3	Rojo	?

Terminología Machine Learning

- Etiqueta

El valor categórico que se asigna a una instancia en clasificación.

Etiqueta 1: Auto.

Etiqueta 2: Moto.

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.

Terminología Machine Learning

- Entrenamiento

Proceso en que el modelo ajusta sus parámetros

Terminología Machine Learning

- Validación

Proceso en que evaluamos el rendimiento del modelo entrenado con un conjunto de datos separado para ajustar hiperparámetros y evitar sobreajuste

Terminología Machine Learning

- Test/prueba

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

Preprocesamiento de datos

- La calidad de los datos impacta en la efectividad del modelo.
- Puede ser conveniente procesarlos antes de entrenar

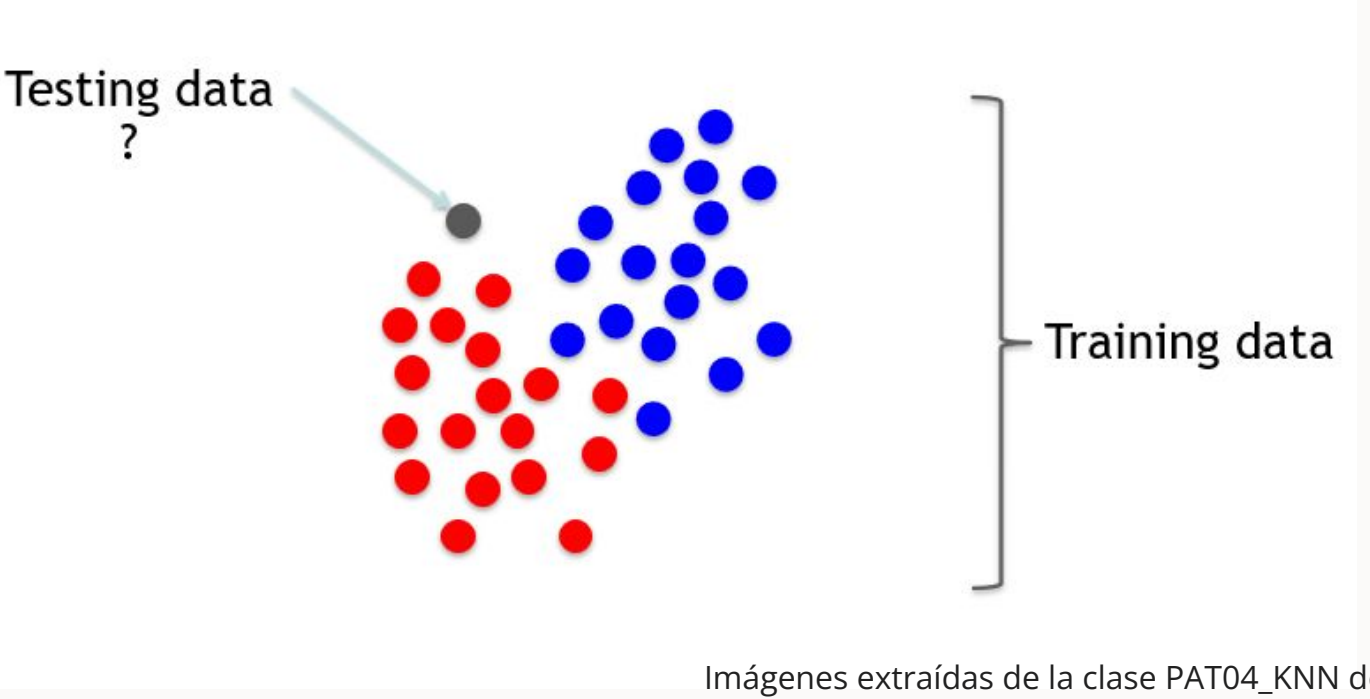
Preprocesamiento de datos

- Codificación de los datos
- Escalamiento y normalización
- Reducción de dimensionalidad
- Imputación

Algoritmo: k-nearest neighbor (KNN)

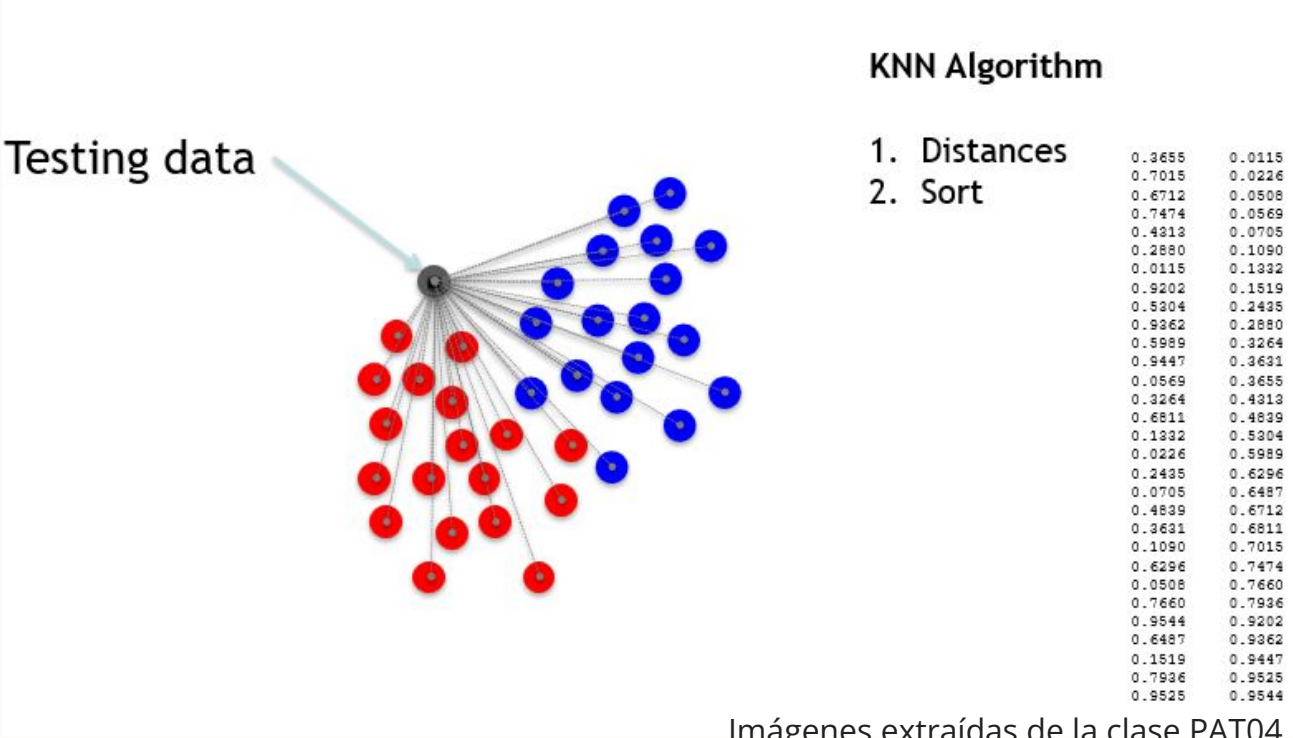
- Clasificación de datos nuevos
- No hay entrenamiento como tal
- De los algoritmos más simples para probar experimentos rápido
- Basado en distancias euclidianas. Preprocesamiento y normalización es fundamental.

Algoritmo: k-nearest neighbor (KNN)



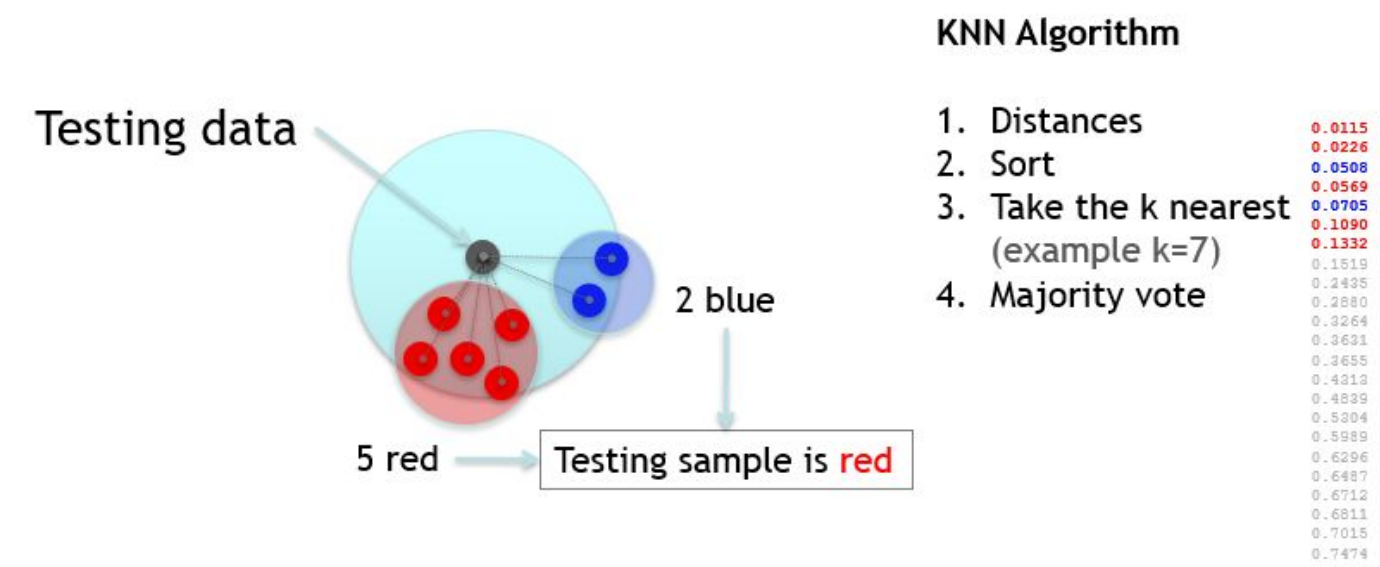
Imágenes extraídas de la clase PAT04_KNN del profesor Domingo Mery

Algoritmo: k-nearest neighbor (KNN)



Imágenes extraídas de la clase PAT04_KNN del profesor Domingo Mery

Algoritmo: k-nearest neighbor (KNN)



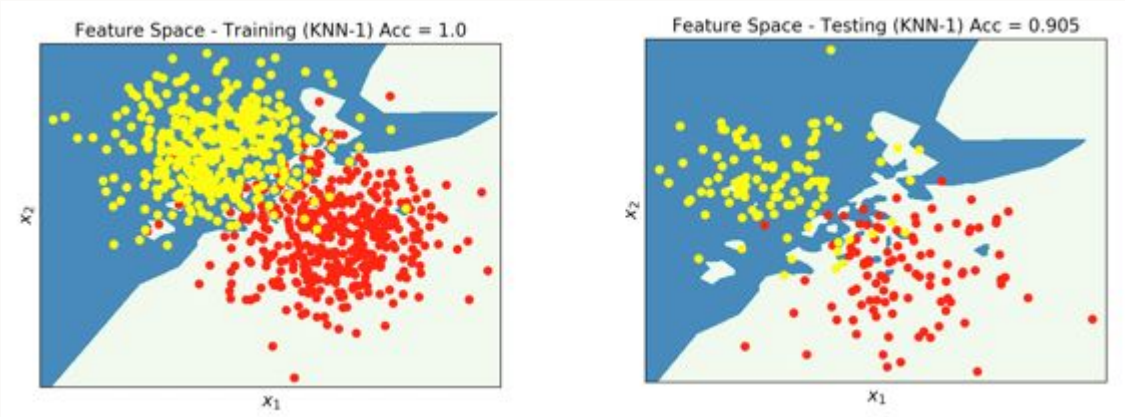
Imágenes extraídas de la clase PAT04_KNN del profesor Domingo Mery

Algoritmo: k-nearest neighbor (KNN)

- ¿Qué parámetros tiene este modelo?
- Y ¿Qué hiper parámetros tiene?

Algoritmo: k-nearest neighbor (KNN)

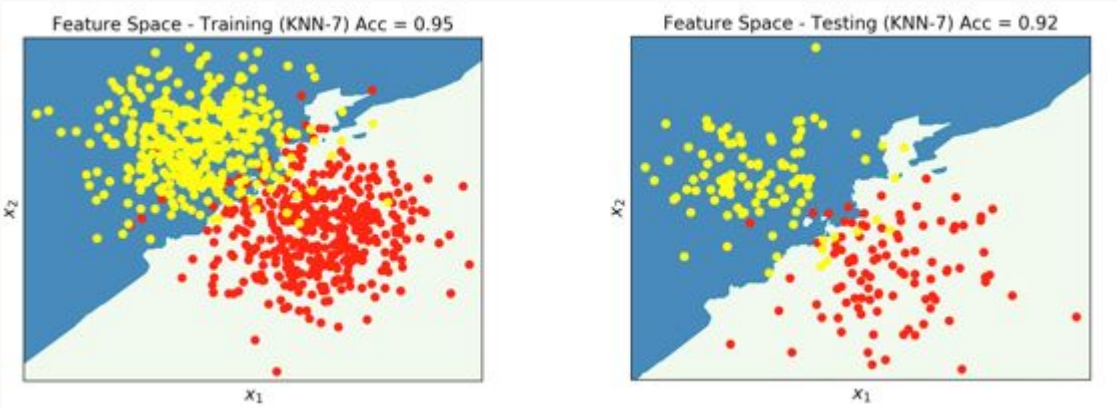
K=1



Imágenes extraídas de la clase PAT04_KNN del profesor Domingo Mery

Algoritmo: k-nearest neighbor (KNN)

K=7



Imágenes extraídas de la clase PAT04_KNN del profesor Domingo Mery

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

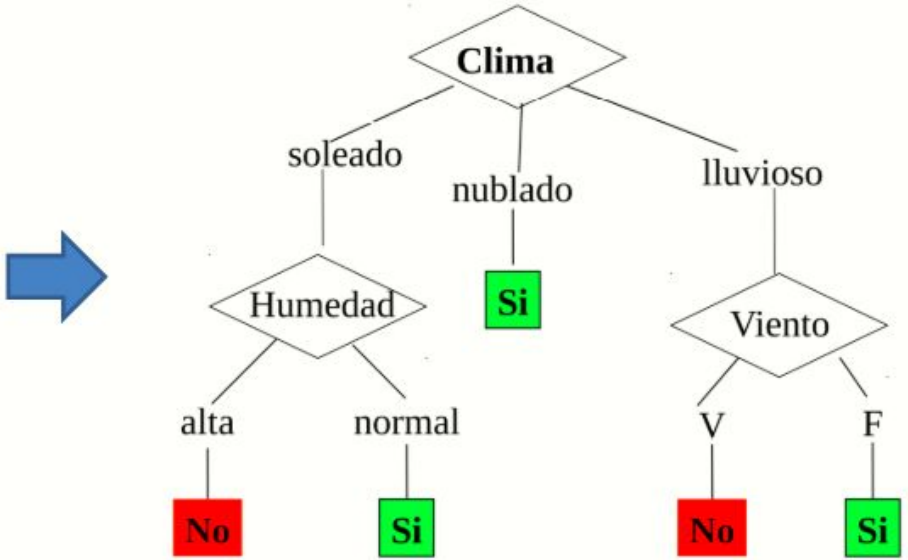


Imagen obtenida de la clase de árboles del profesor Hans Löbel 2023-1

Árboles de decisión

- Modelos Predictivos:
  - Usados para clasificación y regresión.
  - Organizan datos para dividirlos en diferentes clases.

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:
  - Nodos hoja:
    - Cada hoja representa una clase sobre la característica a predecir.
    - Pueden haber varias hojas para una misma clase.

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 métricas 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

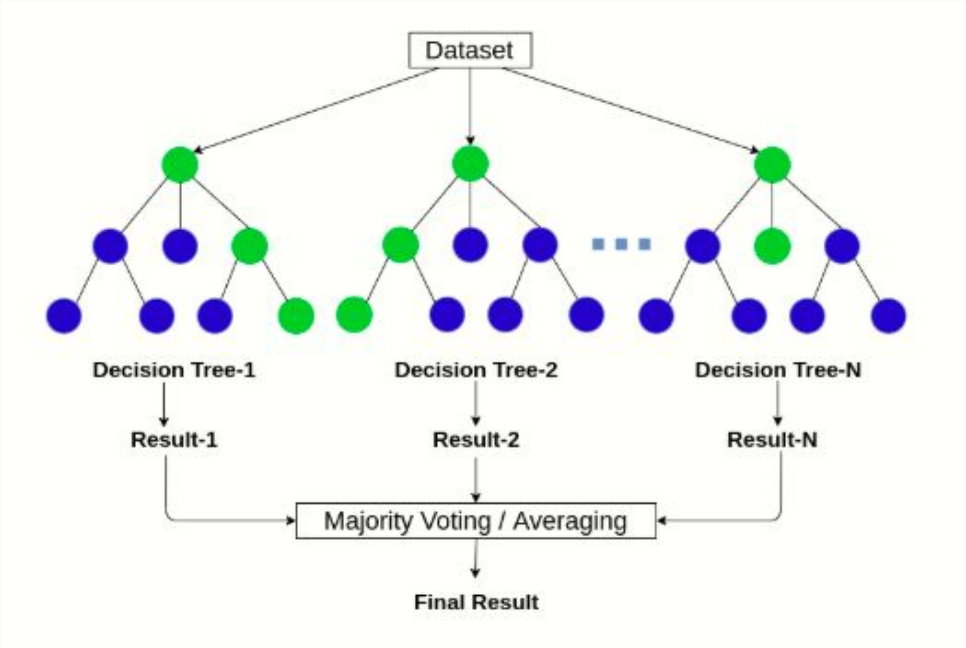


Imagen obtenida de la clase de árboles del profesor Hans Löbel 2023-1

Ensamblajes: Random Forest

- Algoritmo de ensamblaje de tipo *bagging*.
- Un árbol para cada muestra aleatoria (subconjunto) de los datos.
- Árboles de poca profundidad para evitar el overfitting.
- La predicción final se obtiene por mayoría entre las predicciones de los árboles del bosque

Hiperparámetros: RF

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

Ensamblajes: XGBoosting

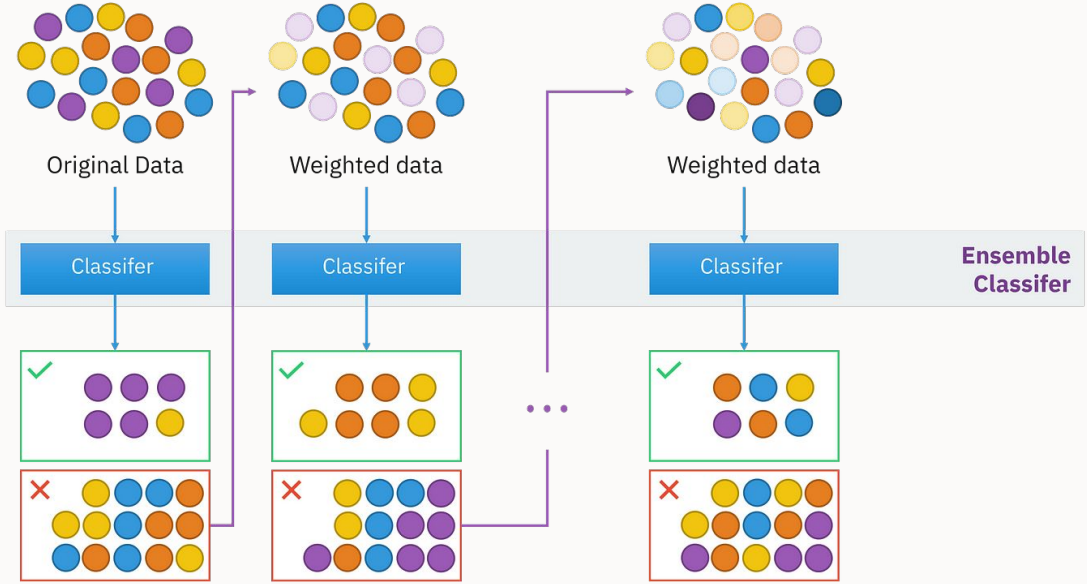


Imagen obtenida de https://medium.com/@chandu.bathula16/machine-learning-concept-53-xgboosting-adaboosting-663cd8c920e2

Ensamblajes: XGBoosting

- Algoritmo de ensamblaje de tipo *boosting*.
- 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
- **subsample:** Proporción de muestras a utilizar para entrenar cada modelo

Ejemplo en código

Ayudantía 8

KNN y árboles

Por Ignacio Villanueva y Kaina Galdames

20 de mayo 2024