Clasificación KNN

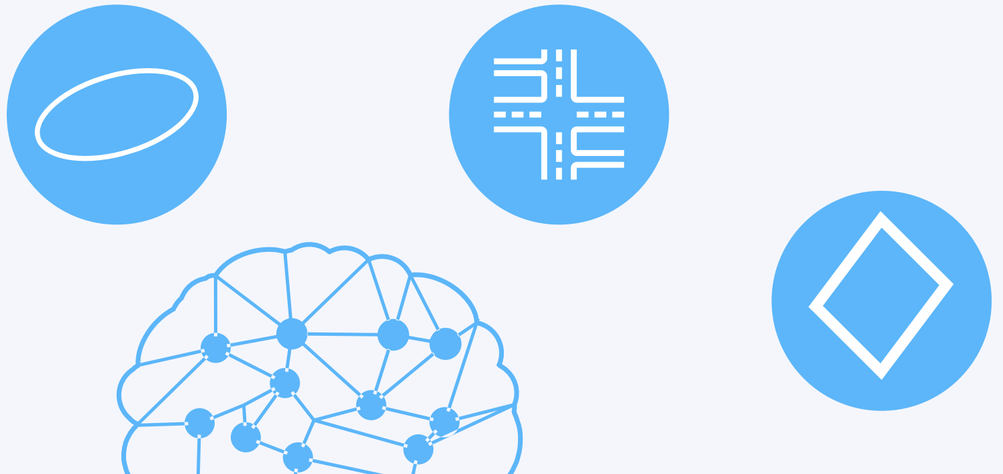
Juan David Bonilla Sotelo
Kevin Andrés Leal Pérez
Stefanía Rojas García

Introducción

Algoritmo no parametrico que opta por recopilar información sobre la estructura y utilizarla como información apriorí para las predicciones.

Funcionamiento

Parte de la idea del algoritmo es calcular distancias con respecto a una clasificación y por tanto encontrar sus vecinos más cercanos mediante el voto de etiquetas, para ello se podría hablar de cuatro tipos de distancia según el entorno



Paso a paso dataiku

01 Importe las bibliotecas de Python relevantes

02 Importar los datos

03 Leer/limpiar/ajustar los datos (si es necesario)

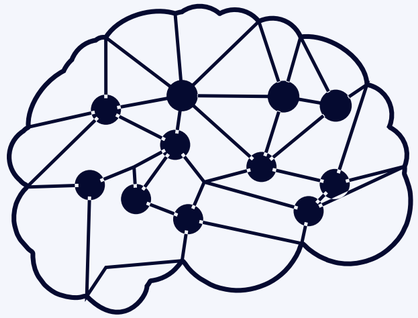
04 Crear una división de entrenamiento/prueba

05 Crear el objeto modelo kNN

06 Ajuste el modelo

07 Predecir

08 Evaluar la precisión

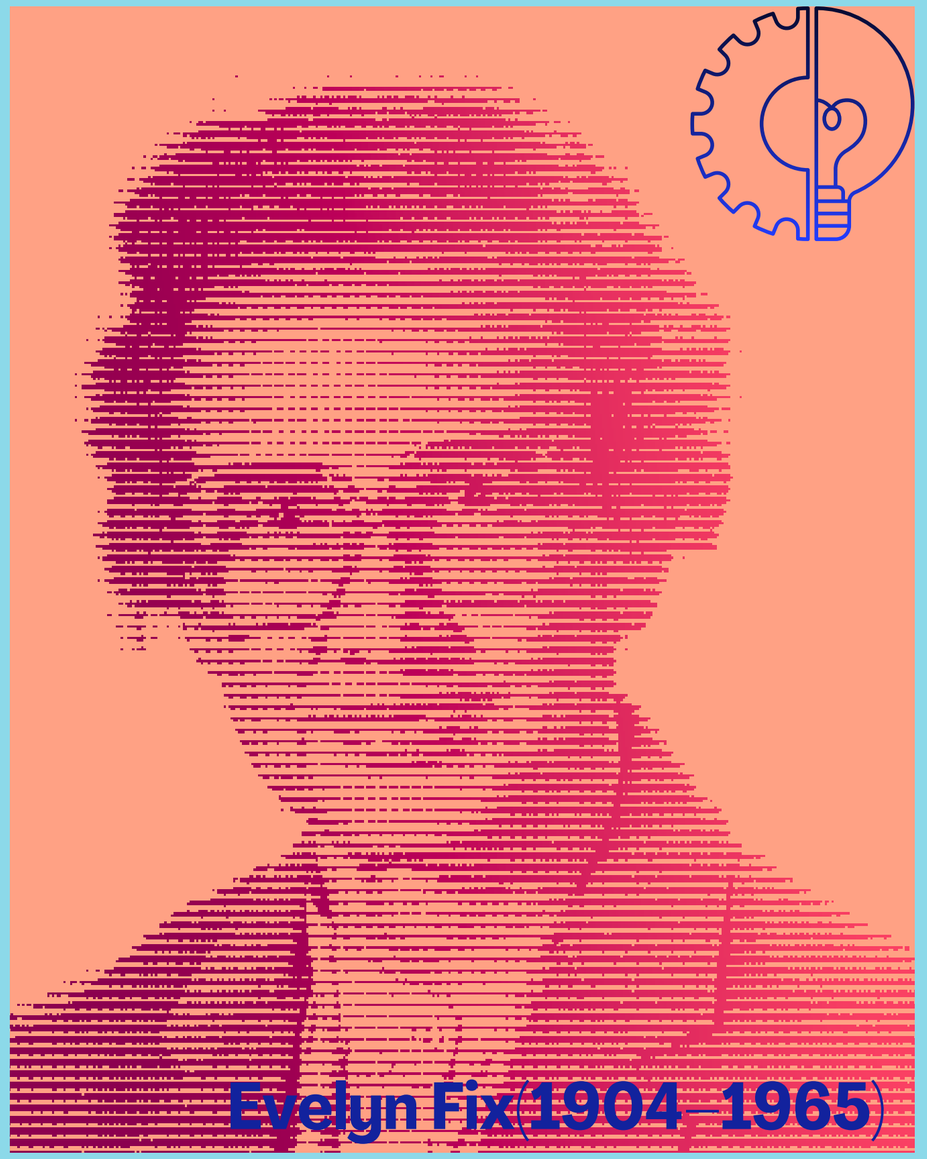


Nota histórica

informe de análisis técnico en 1951

Estadísticos de Berkeley que introdujeron un método de clasificación no paramétrico (análisis discriminante)

"Estos límites son los más estrechos posibles, para todas las distribuciones subyacentes adecuadamente uniformes" (1967,P.Hart,p1).

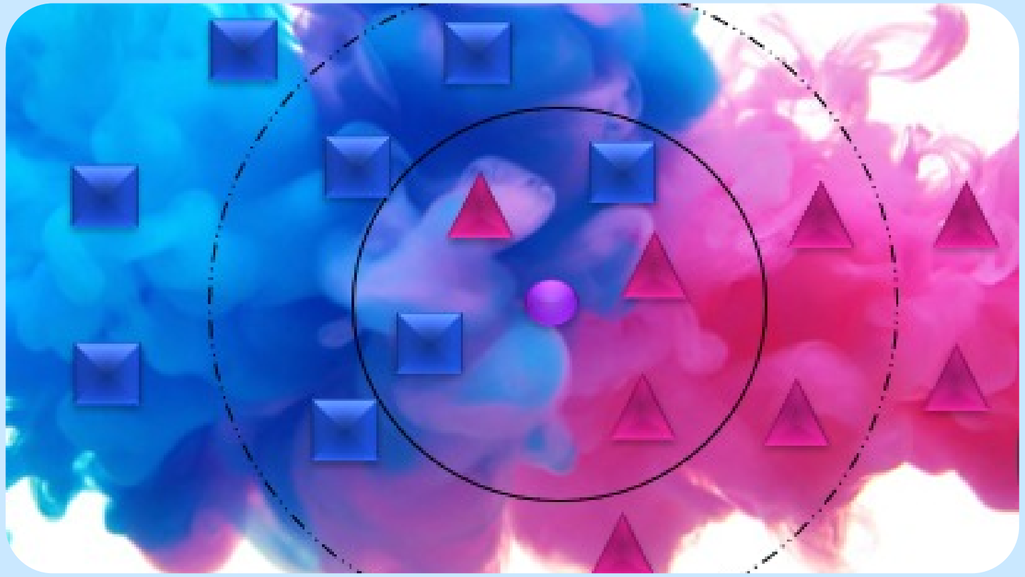


Evelyn Fix(1904–1965)

Distancias clásicas

Euclidiana

Mahalanobis

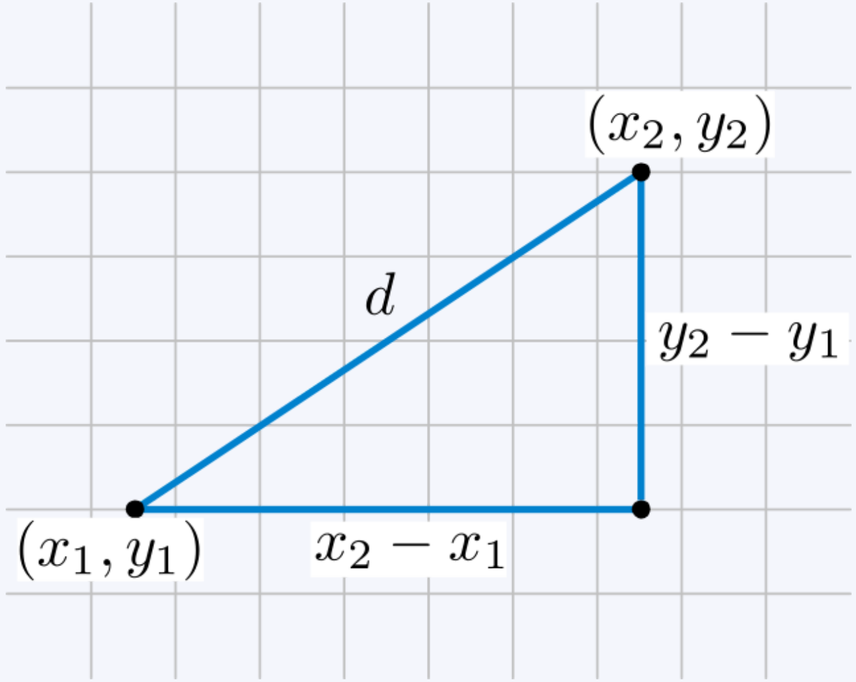


Manhattan

Minkowski

Distancia Euclidiana

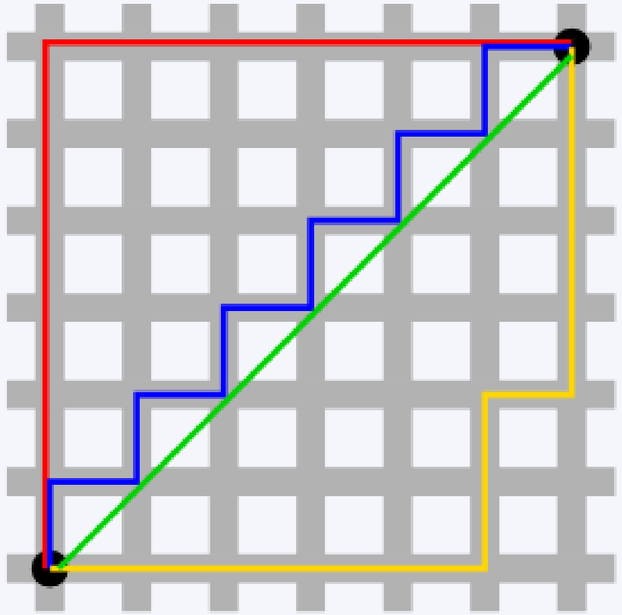
- Se puede deducir del teorema de pitagoras
- Puntos en un plano bidimensional



$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Distancia Manhattan

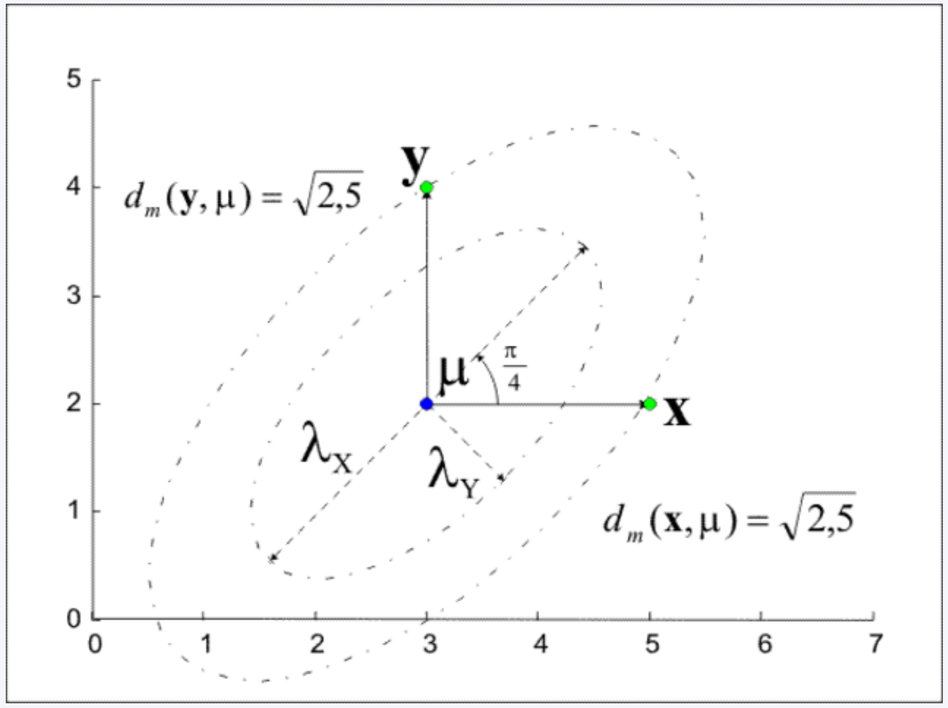
- Suma de las diferencias absolutas
- Cuadrícula, que muestra como ir en diferentes direcciones



Distancia de Manhattan (líneas amarilla, rojas y azules) y distancia Euclidiana (línea verde)

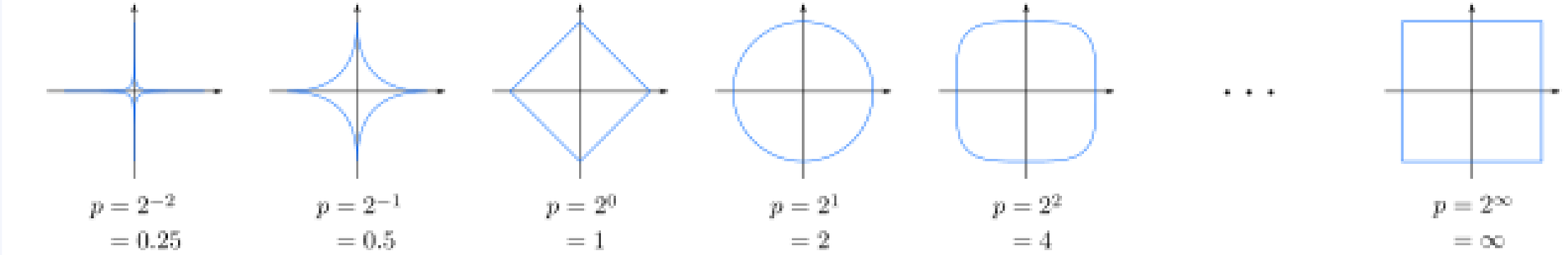
Distancia Mahalanobis

- Puntos en un espacio multivariado
- Entre más distancia de Mahalanobis, más lejos está el dato del centroide



Distancia Minkowski

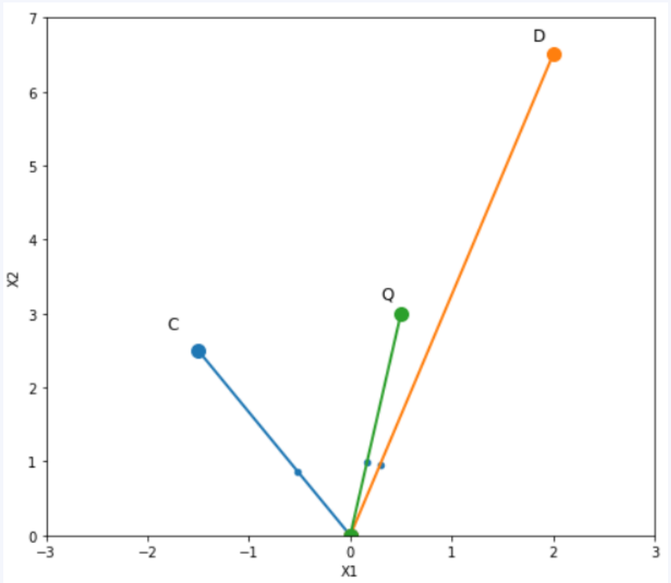
- Generalización de las distancias Manhattan y Euclidiana
- Se agrega un parámetro ***p***



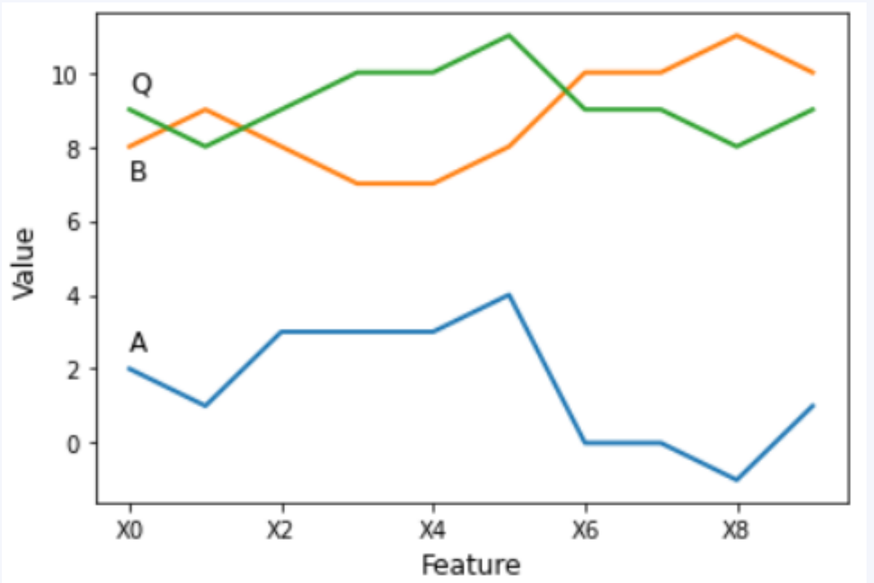
Si el parámetro p es 1, entonces esta distancia es igual a la de Manhattan y si es 2 es igual a la Euclidiana.

Medidas de Similitud

Coseno



Correlación



Pseudo-distancias útiles para datos de otras naturalezas como texto o comparaciones en asignación de recursos

Implementación en Python

Caso 1: Cáncer de mama Wisconsin (Diagnóstico)

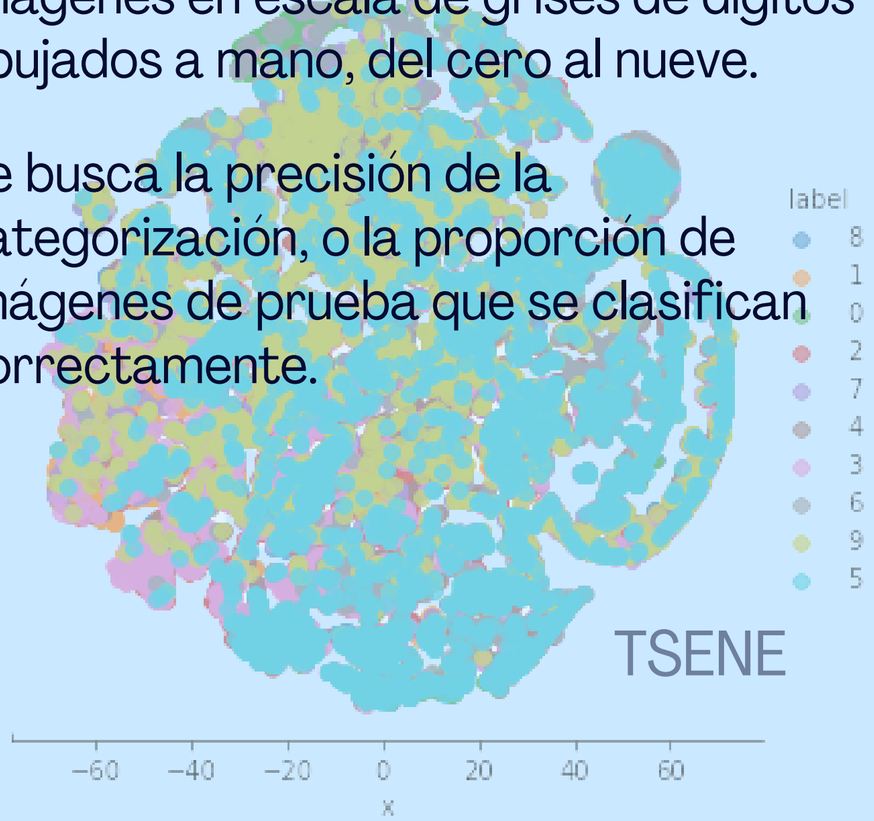
Profesores de la Universidad de Wisconsin en el año 1995

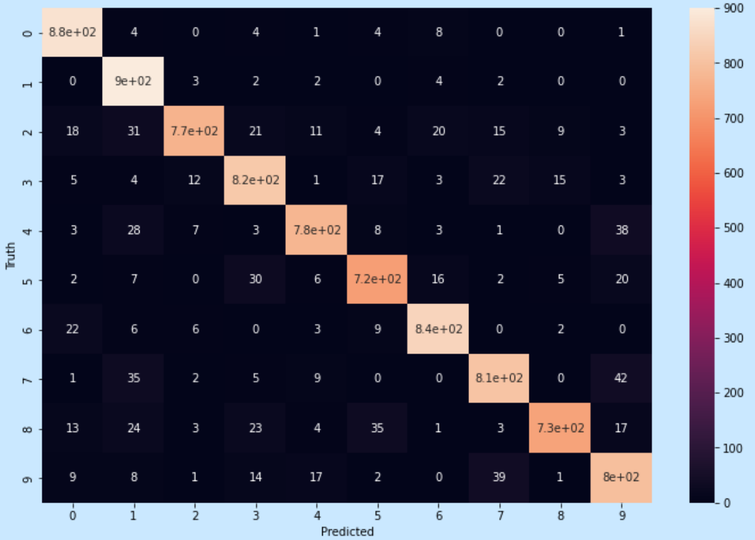
Predecir si los tumores encontrados eran benignos o malignos

Análisis completo en: https://github.com/StefaniaRojas/Mineria-de-datos/blob/main/Trabajo%20Miner%C3%ADa/Trabajo_Miner%C3%ADa.ipynb

Caso 2: Reconocedor de dígitos

- imágenes en escala de grises de dígitos dibujados a mano, del cero al nueve.
- Se busca la precisión de la categorización, o la proporción de imágenes de prueba que se clasifican correctamente.





Análisis completo en: https://github.com/StefaniaRojas/Mineria-de-datos/blob/main/Trabajo%20Miner%C3%ADa/Trabajo_Miner%C3%ADa.ipynb

Caso 3: Clasificación de las descripciones de empresas brasileñas

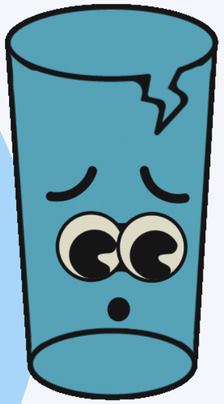
Precisión:

- Euclidiana → 86%
- Coseno → 92%

Caso 4: Clasificación de los tipos de vidrio

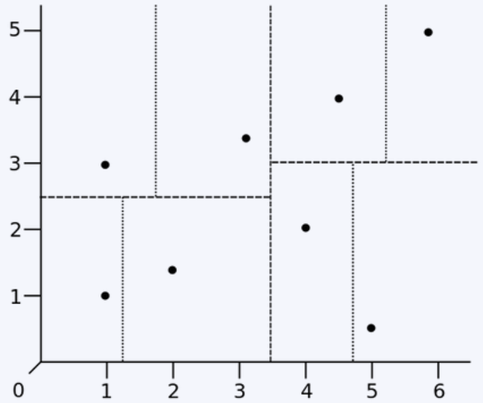
Precisión:

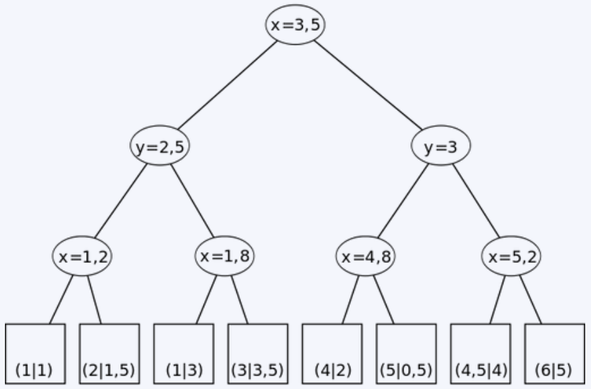
- Euclidiana → 65%
- Correlación → 67%

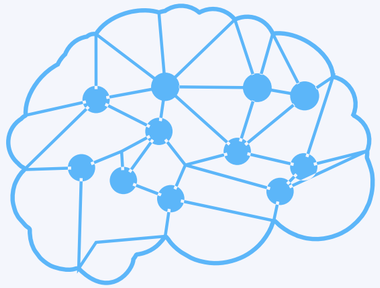


Algoritmos de Optimización

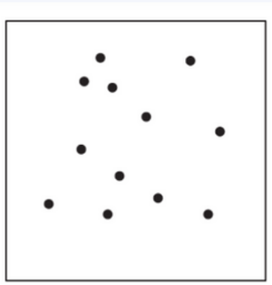
Kd-tree



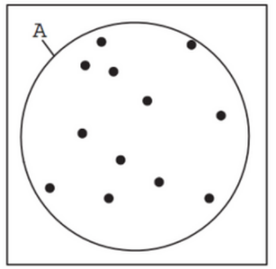




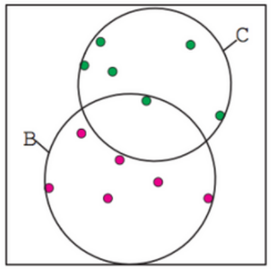
Ball-tree



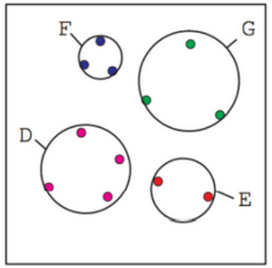
1a. A dataset



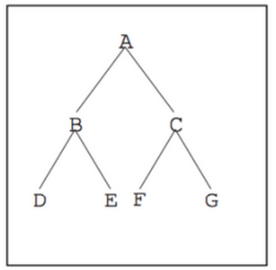
1b. Root node



1c. The 2 children



1d. The 4 grandchildren



1e. The internal tree structure

Comparaciones

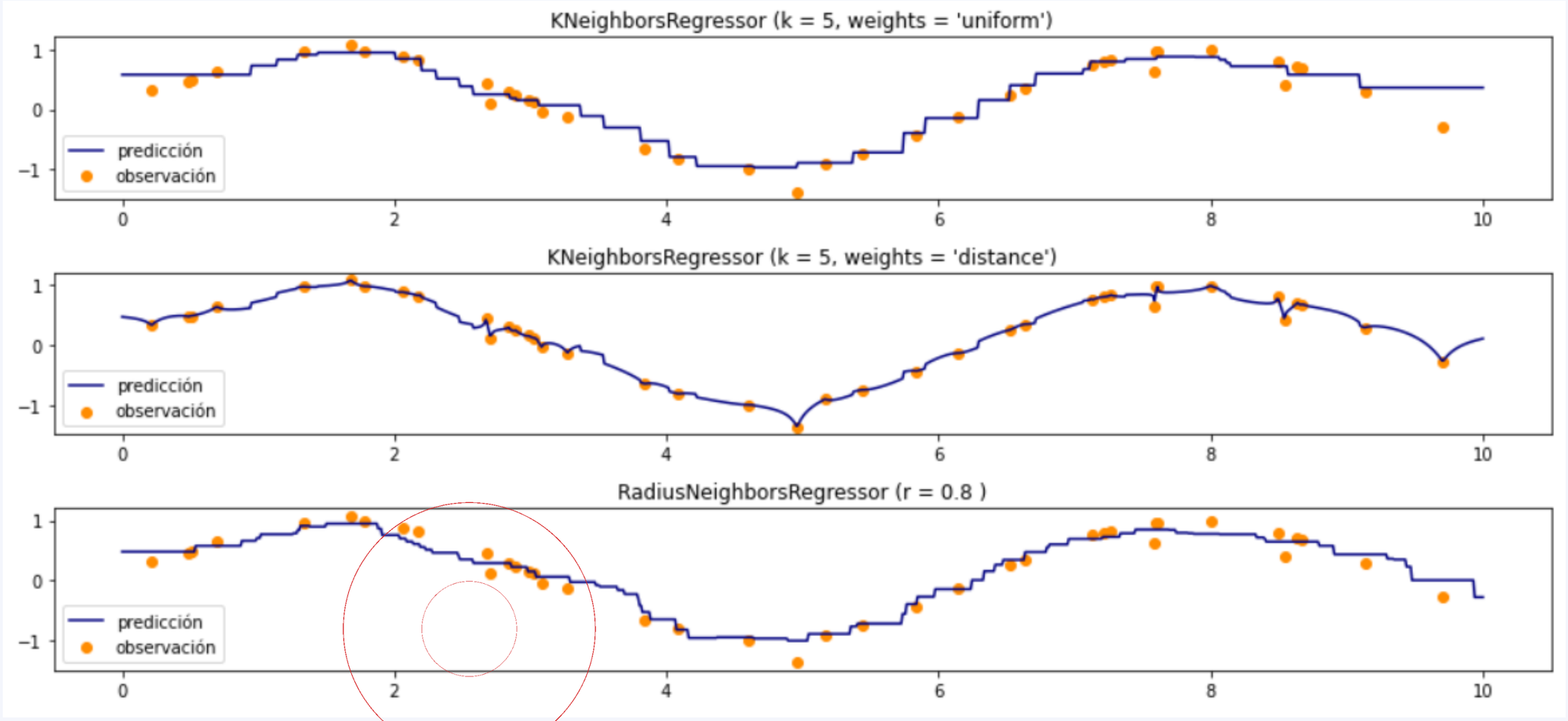
50.000 muestras
9 variables

- Fuerza Bruta: 59.9 seg
- Kd-tree: 22.4 seg

25.000 muestras
25 variables

- Fuerza Bruta: 16.4seg
- Kd-tree: 64 seg
- Ball-tree: 49.6 seg

REGRESIÓN KNN



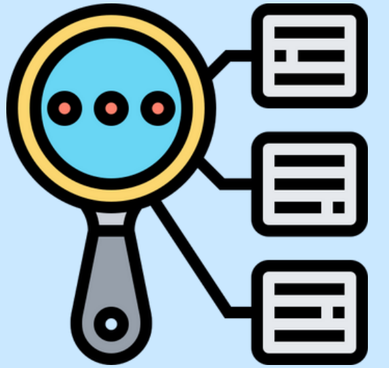
Ventajas

- No parámetrico

- Algoritmo simple

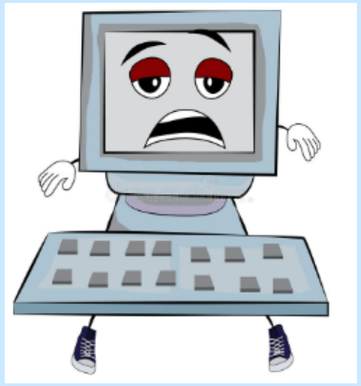
- Precisión alta

- No es sensible a valores atípicos

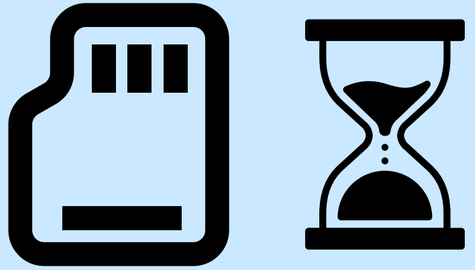


Desventajas

- Coste computacional

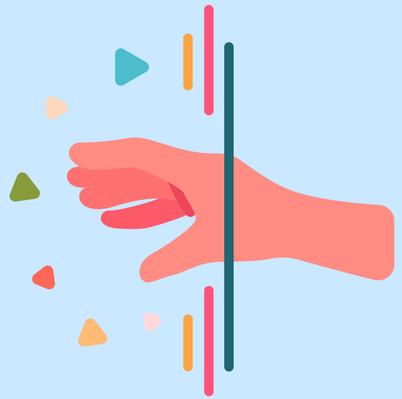


- Memoria y tiempo



- No es tan sencillo hacer clasificación para muchas variables

- Maldición de la dimensionalidad



Link del Repositorio:

https://github.com/StefaniaRojas/Mineria-de-datos/tree/main/Trabajo%20Miner%C3%ADa