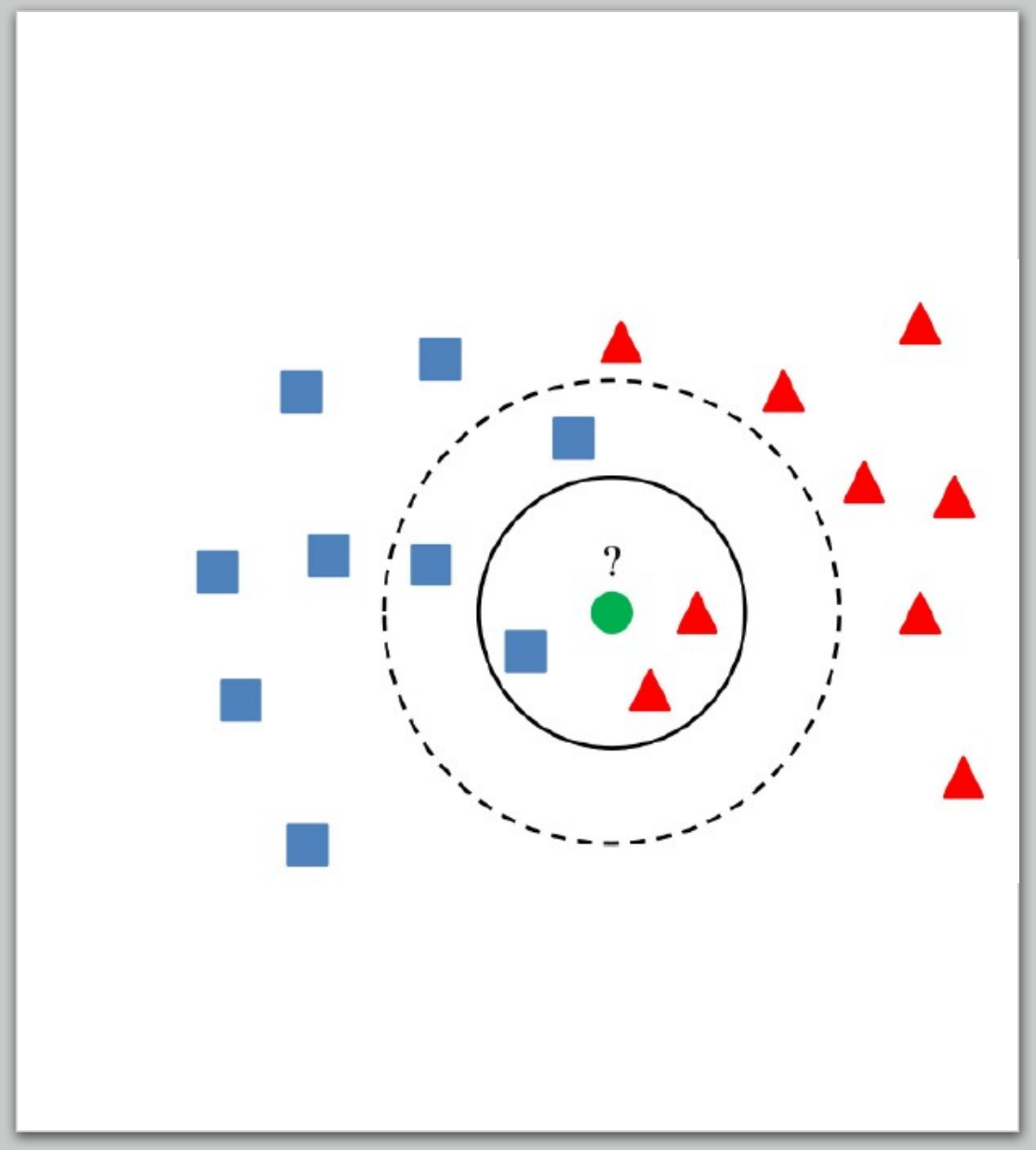
Machine Learning
KNN

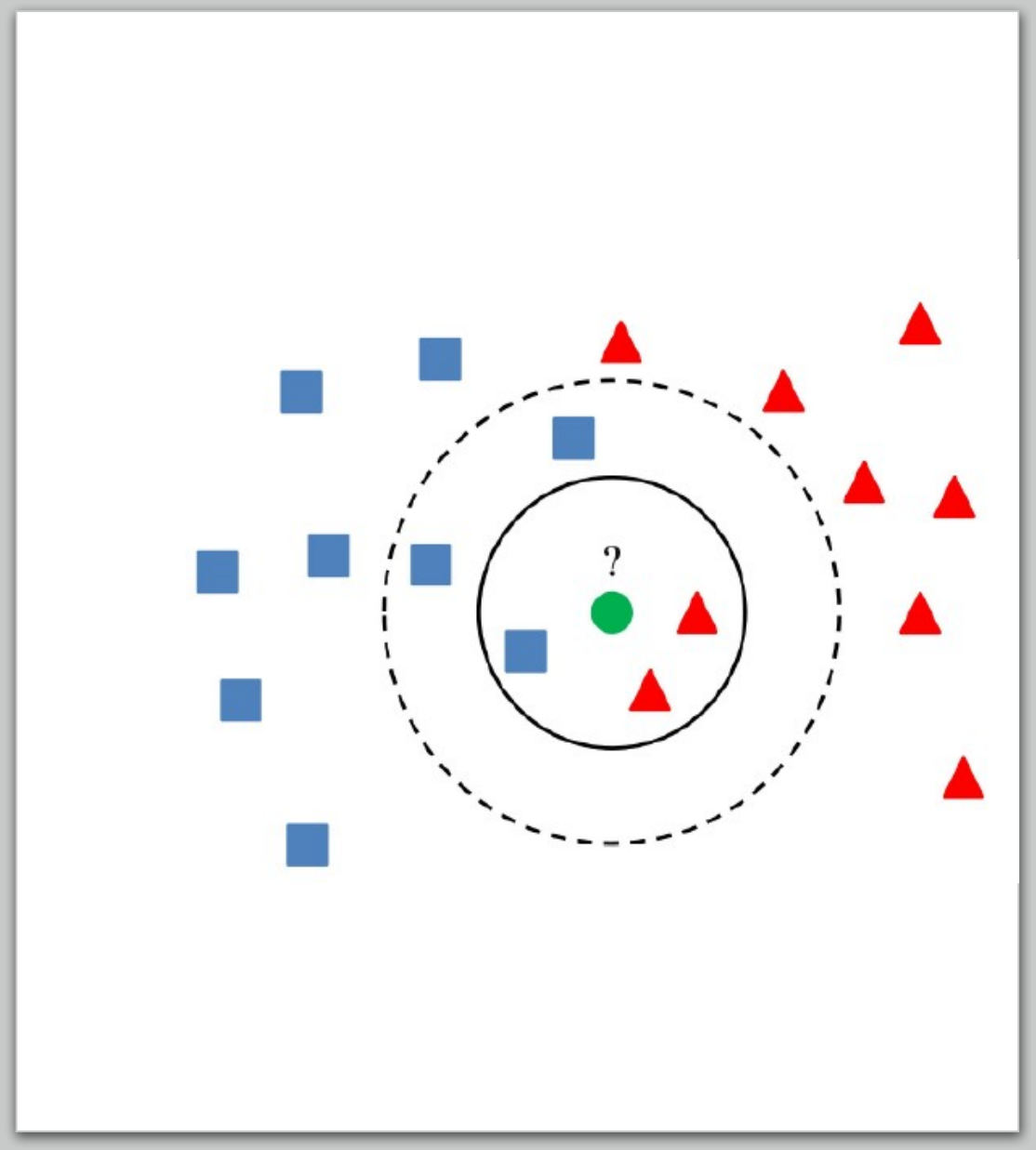
¿Qué es el KNN?

- *KNN por sus siglas del inglés k-nearest neighbor.*
- El método KNN es un método de **clasificación** supervisado que estima la probabilidad a posteriori de que un elemento x pertenezca a la clase Cj a partir de la información proporcionada por el conjunto seleccionado.
- La clasificación se calcula a partir de un voto de mayoría simple de los vecinos más cercanos de cada punto: a un punto de consulta se le asigna la clase de datos que tiene más representantes dentro de los vecinos más cercanos del punto.



¿Cómo funciona?

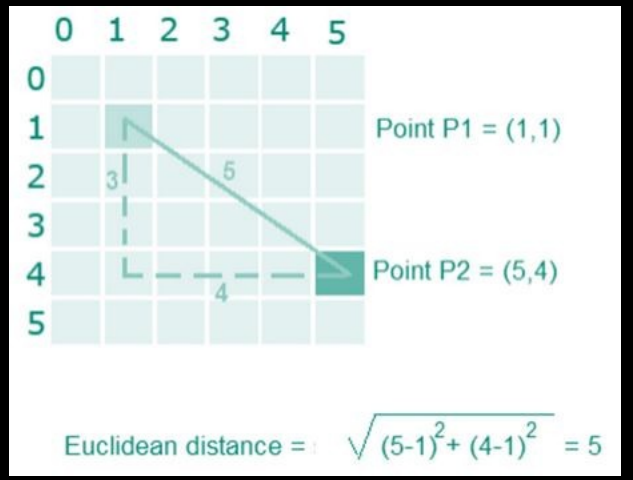
- ¿Qué criterio utilizamos? Normalmente la distancia Euclídea.
- Sklearn, utiliza el aprendizaje basado en los K vecinos más cercanos de cada punto de consulta, donde K es un valor entero especificado por el usuario.
- Se suelen utilizar números impares para evitar empates.
- Lazy Learn Algorithm: no entrena y simplemente guarda el dataset. La computación se produce en predicción.
- Funciona bien con datasets pequeños
- No requiere de suposiciones en los datos
- Recomendable escalar



Distancia Euclídea

Euclidean Distance between (x_1, y_1) and(x_2, y_2) = $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

- El enfoque euclidiano, es la medida de distancia más utilizada para calcular la distancia entre las muestras de prueba y los valores de datos entrenados.
- Podemos calcular la distancia euclídea en un espacio n-dimensional.

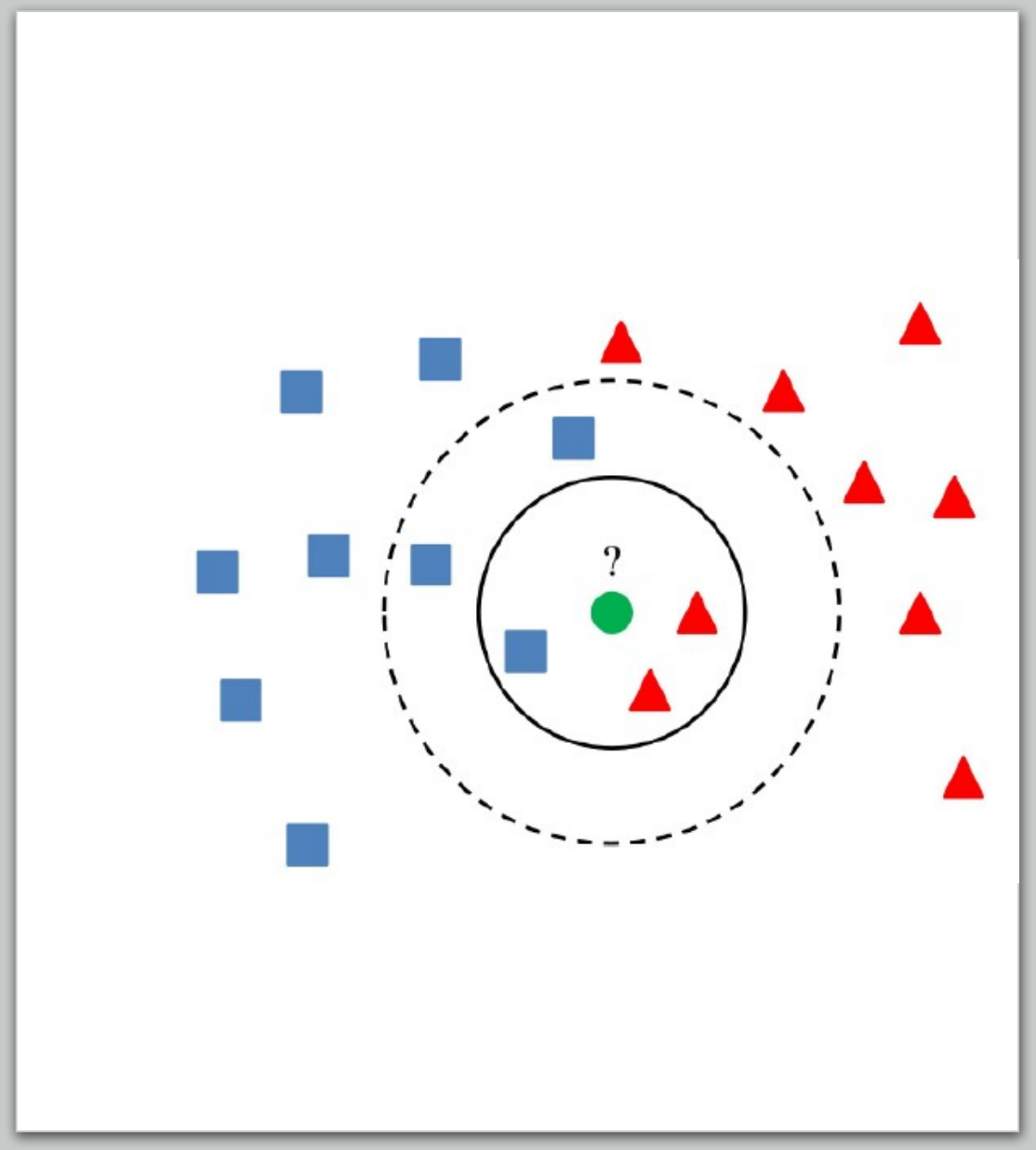


En general, la distancia euclidiana entre los puntos $P = (p_1, p_2, \ldots, p_n)$ y $Q = (q_1, q_2, \ldots, q_n)$, del espacio euclídeo *n*-dimensional, se define como:

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

¿Cómo seleccionamos el valor óptimo de K?

- La mejor elección de k depende fundamentalmente de los datos.
- No existen métodos estadísticos predefinidos para encontrar el valor más favorable de K.
- Iniciamos un valor K aleatorio y calculamos.
- Investigación “iterativa”.
- Elegir valores pequeños de K conduce a límites de decisión muy sensibles o inestables.
- Para clasificar, elegir valores mayores de K nos suaviza esos límites de decisión (reducimos efecto ruido).



Escalar variables

- Los algoritmos basados en la distancia se ven afectados por la escala de las variables.
- Por ello, debemos tener las variables en rangos similares.
- No tiene sentido calcular distancias para variables con escalas muy diversas: variable edad vs variable ingresos.
- Para ello utilizaremos los distintos tipos de escalado:
  - Estandarización
  - Min-Max scaler
  - etc

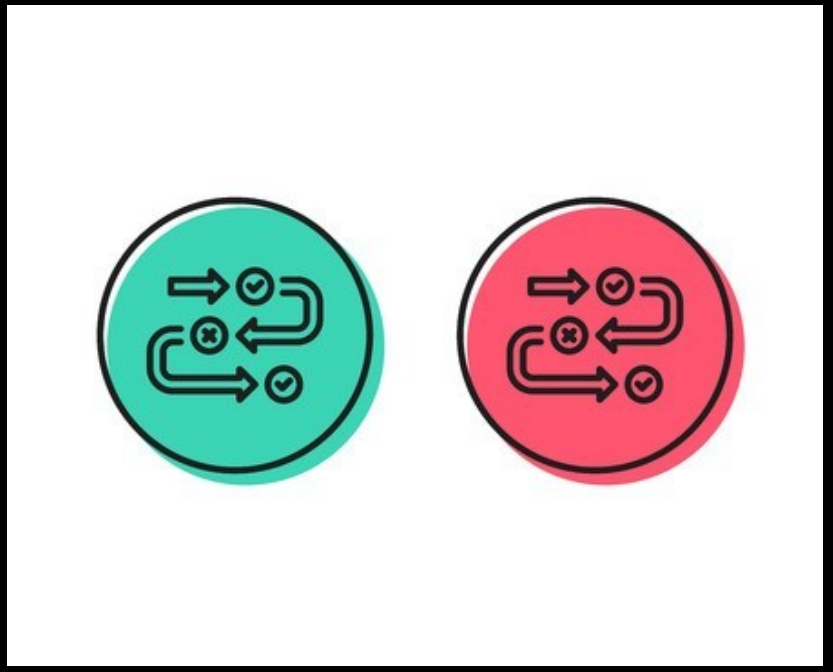
Distancia euclídia = [(100000–80000) ^ 2 + (30–25) ^ 2] ^ (1/2)

¿Cómo evaluamos el modelo?

- Cambiar el valor de K puede afectar el rendimiento del modelo, por lo que debemos buscar su valor óptimo.

- Para escoger el valor óptimo de K utilizaremos las métricas de medida de performance:
    - Precisión
    - F1-score
    - Matriz de confusión
    - AUC
    - etc

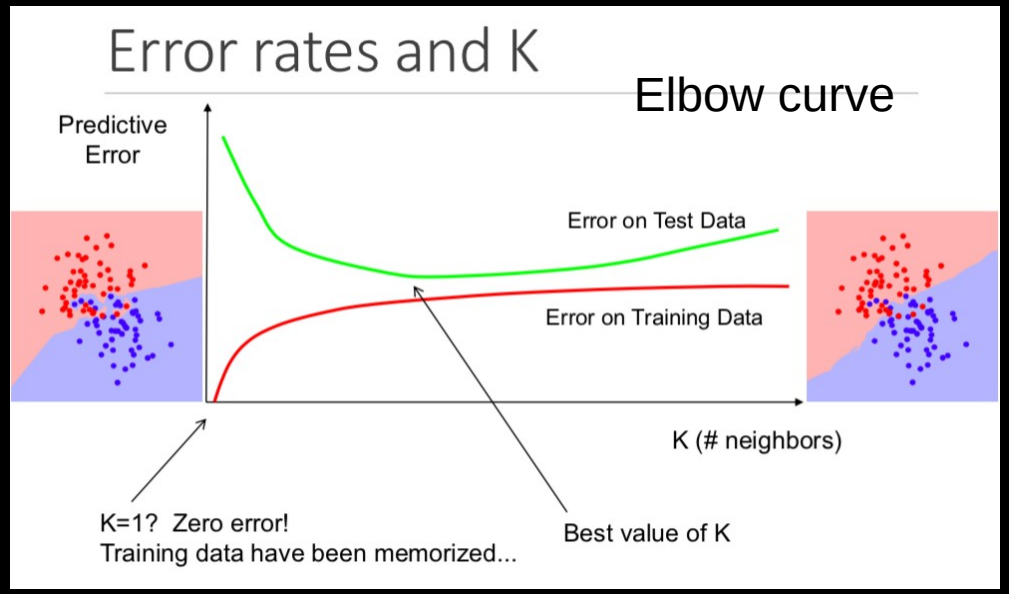
- Problemas de rendimiento para datasets muy grandes.

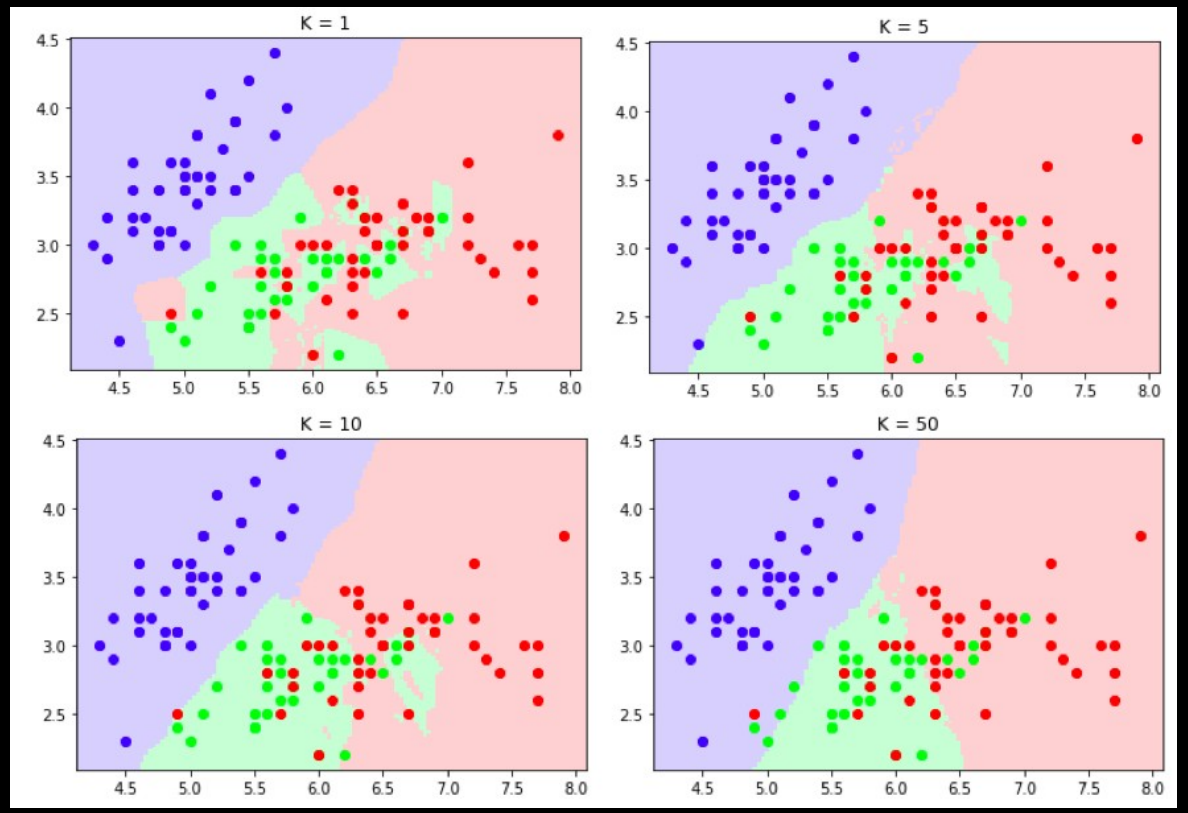


Overfitting en KNN

‘K’s muy bajos producen overfitting

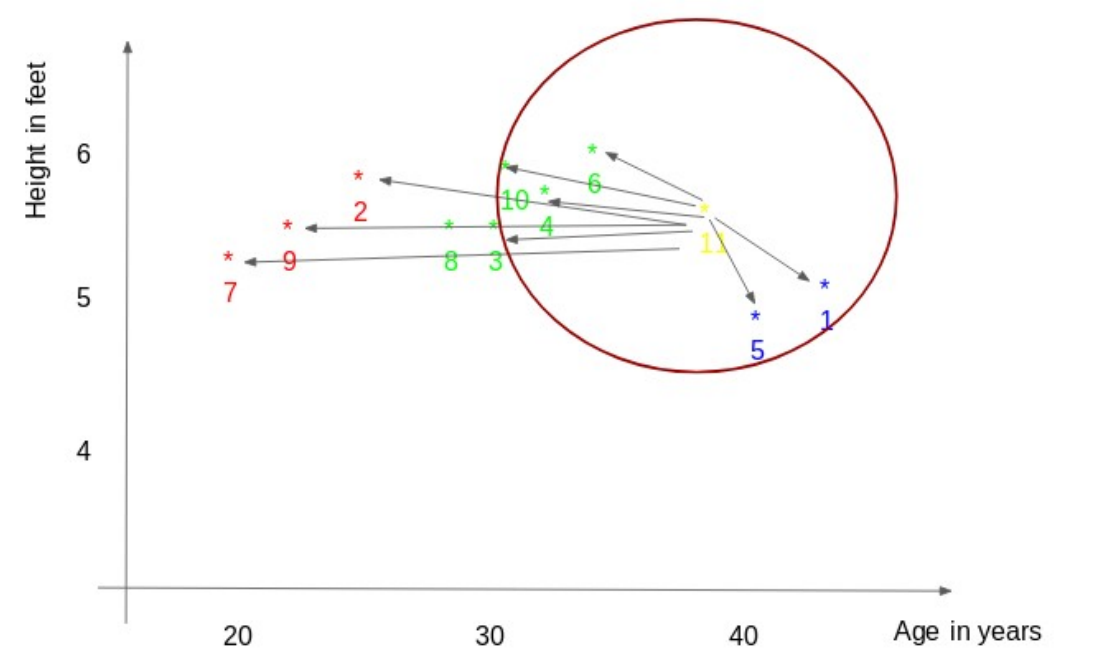
‘K’s altos son más consistentes en los rangos de separación





KNN Regressor

- Se calculan las distancias de cada punto al target
- Se eligen K-vecinos
- El output será la media de los targets de los K-vecinos



ID	Height	Age	Weight
1	5	45	77
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
10	5.6	32	58

(77+59+72+60+58)/5

Preguntas