DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de Datos

Regresión y kNN

2do Cuatrimestre 2024

Laje, LópezRosenfeld, de Erausquin

¿Cuánto medirá de grande?

Información

Es varón

La mamá es bajita, mide 156

¿Cuánto medirá de grande?

- Sin información → ¿Qué podemos decir?

Estimamos:

¿Cuánto medirá de grande?

- Sin información → ¿Qué podemos decir?

  Necesitamos **datos**

¿Cuánto medirá de grande?

¿Promediamos?

Estimamos: 171.5

¿Cuánto medirá de grande?

+ Sin información ✅
+ Es varón →

Completemos
columna “sexo”

¿Cuánto medirá de grande?

¿Promediamos entre varones?

ESTIMAMOS: 178

¿Cuánto medirá de grande?

- Sin información → ✅
- Es varón → ✅
- Es varón y la mamá bajita →

Completemos columna **“contextura mamá”**

¿Cuánto medirá de grande?

¿Promediamos entre varones de mamás bajitas?

ESTIMAMOS: 173

¿Cuánto medirá de grande?

- Sin información → ✅
- Es varón → ✅
- Es varón y la mamá bajita →✅
- Es varón y la mamá mide 156 →

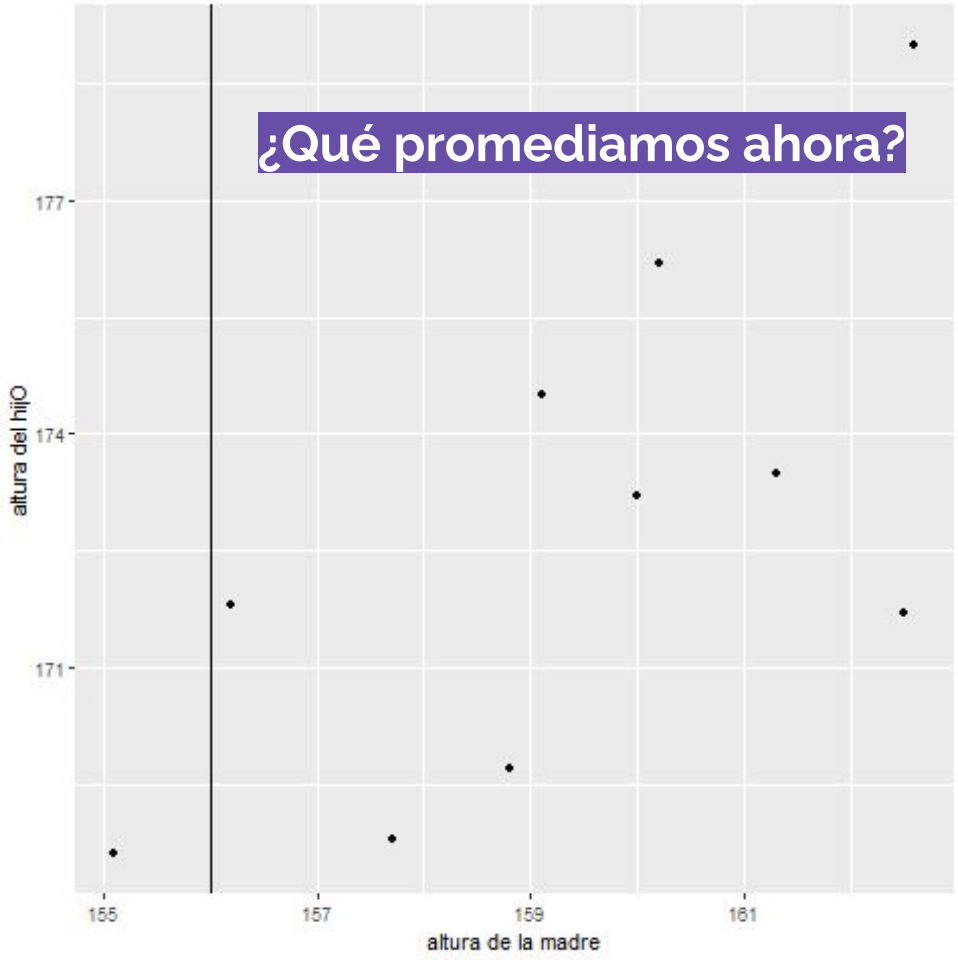
Completemos columna **“altura mamá”**

¿Cuánto medirá de grande?

¿Qué promediamos ahora?

Regresión: estimar la relación entre una variable dependiente (altura cuando sea grande) y una o más variables independientes (sexo, altura de la madre, etc)

¿Qué promediamos ahora?



Una posibilidad: kNN

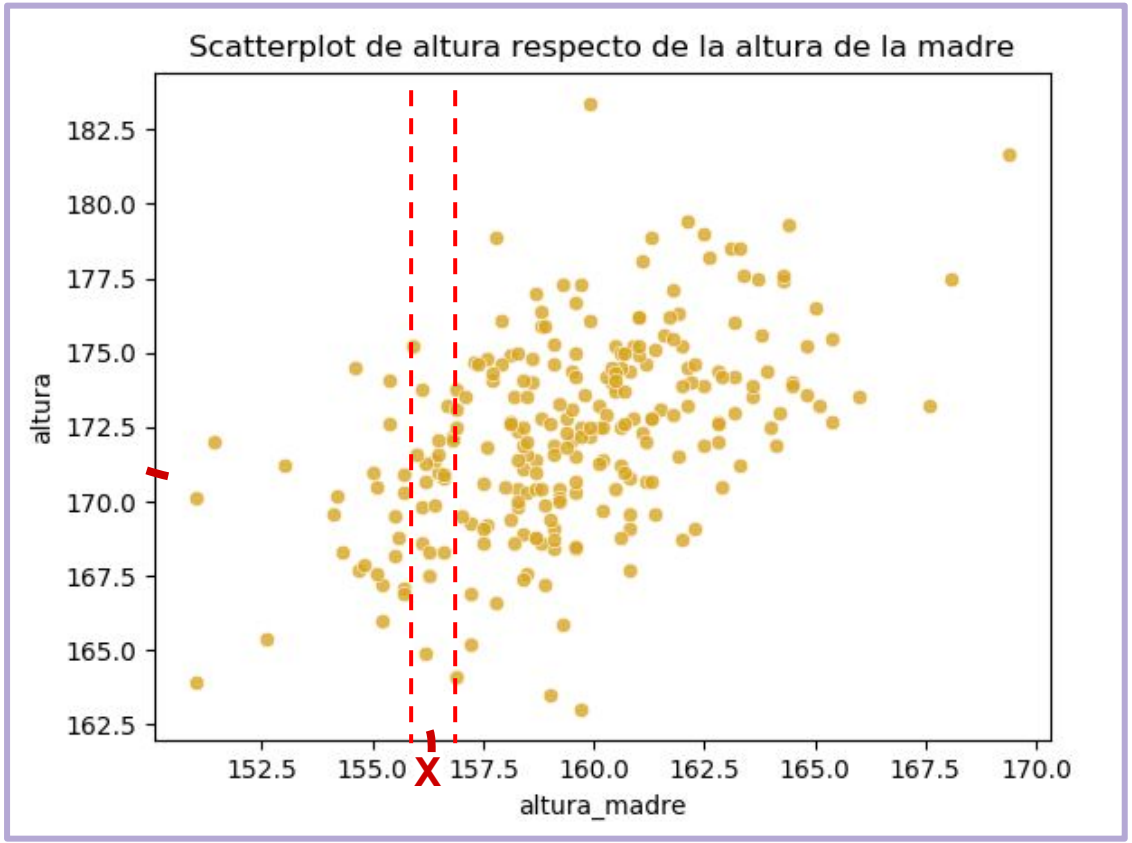
Idea: Promediamos los valores de casos parecidos

kNN: k nearest neighbors - k vecinos más cercanos

Consideramos los k valores más *cercanos** al valor nuevo (altura madre) y promediamos las alturas de esos k varones

*Cercanos: en la o las variables explicativas, y con la distancia que consideremos.

Modelo de kNN



KNN con sklearn

Modelo kNN

+ Modelo no paramétrico
+ Intuitivo y simple
+ Versátil, se adapta a datos que no vienen de una función conocida
+ Al variar el k:
  al bajar el k se adapta más a los datos, genera más variabilidad
  al subir el k se suaviza, resulta más estable

Clasificación con K Nearest Neighbors (KNN)

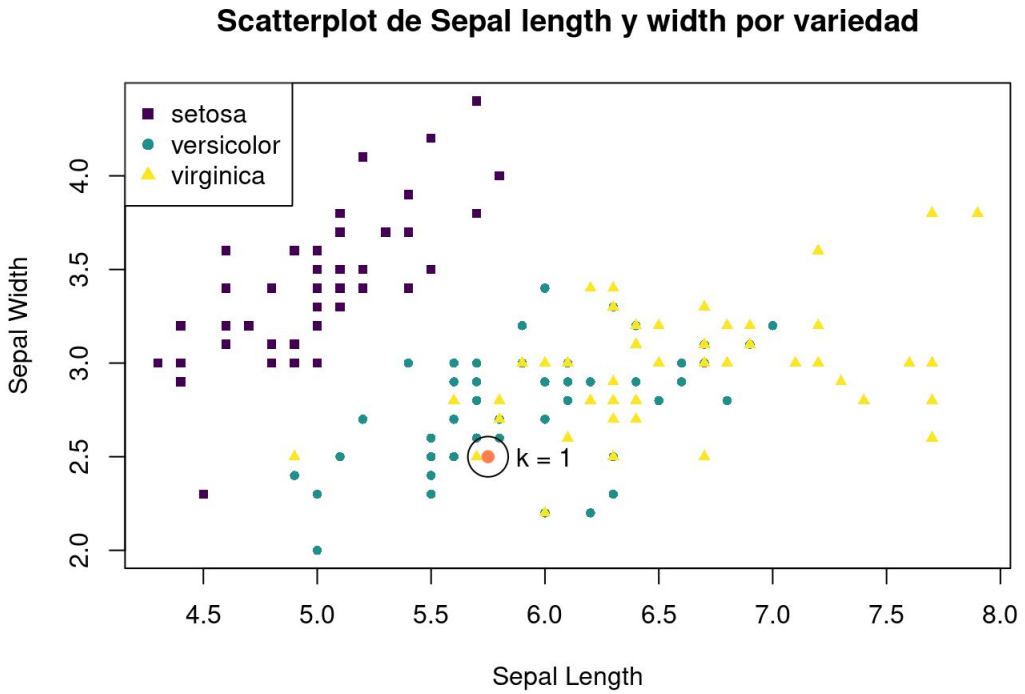
Es similar a cuando lo usamos para regresión:

- Definir una distancia en los atributos
- Buscar los k vecinos más cercanos (según esta distancia)
- Ver qué clases tienen
- Elegir la mayoritaria

Por ejemplo, tomando la distancia euclídea en los atributos Sepal Length y Sepal Width

Buscamos el vecino más cercano, entre los que ya tenemos etiquetados.

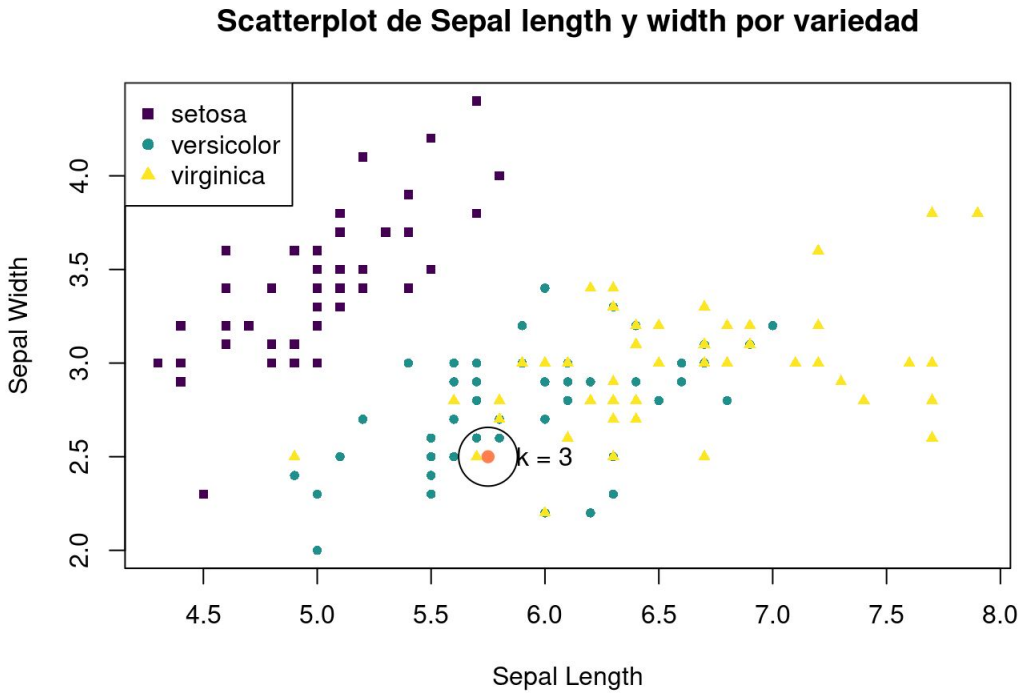
Nos copiamos esa etiqueta.



Por ejemplo, tomando la distancia euclídea en los atributos Sepal Length y Sepal Width

Buscamos los 3 más cercanos, entre los que ya tenemos etiquetados.

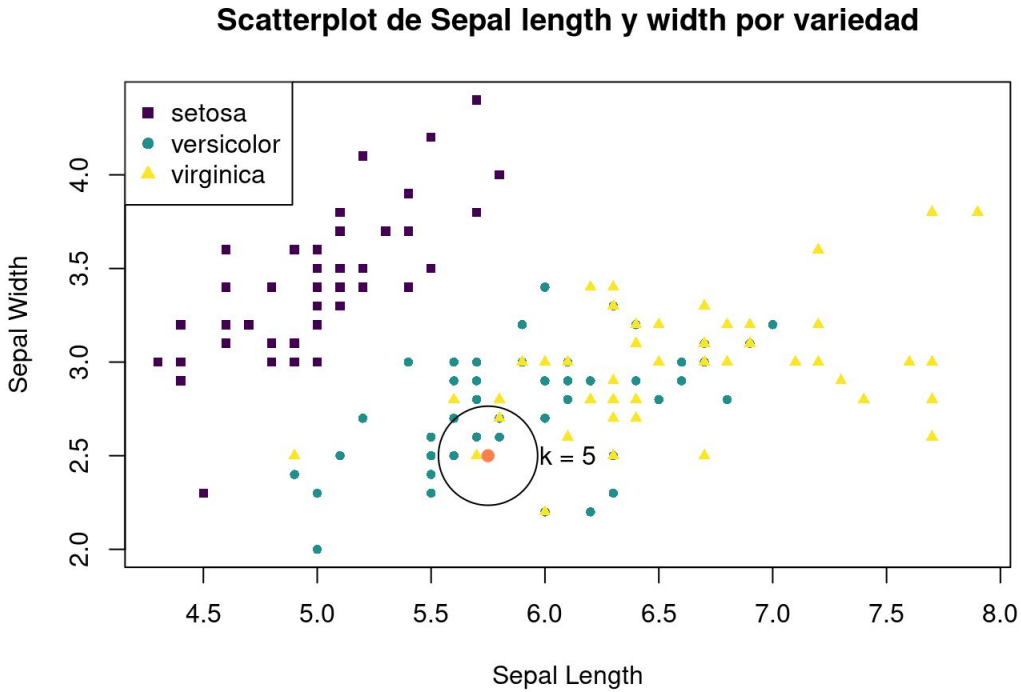
Tomamos la clase mayoritaria.



Por ejemplo, tomando la distancia euclídea en los atributos Sepal Length y Sepal Width

Buscamos los 5 más cercanos, entre los que ya tenemos etiquetados.

Tomamos la clase mayoritaria.



Ejemplos con Iris

Ejemplos, usando todo el dataset, con distintos valores de k.

Vamos a usar los 4 atributos (4 primeras columnas del dataframe).

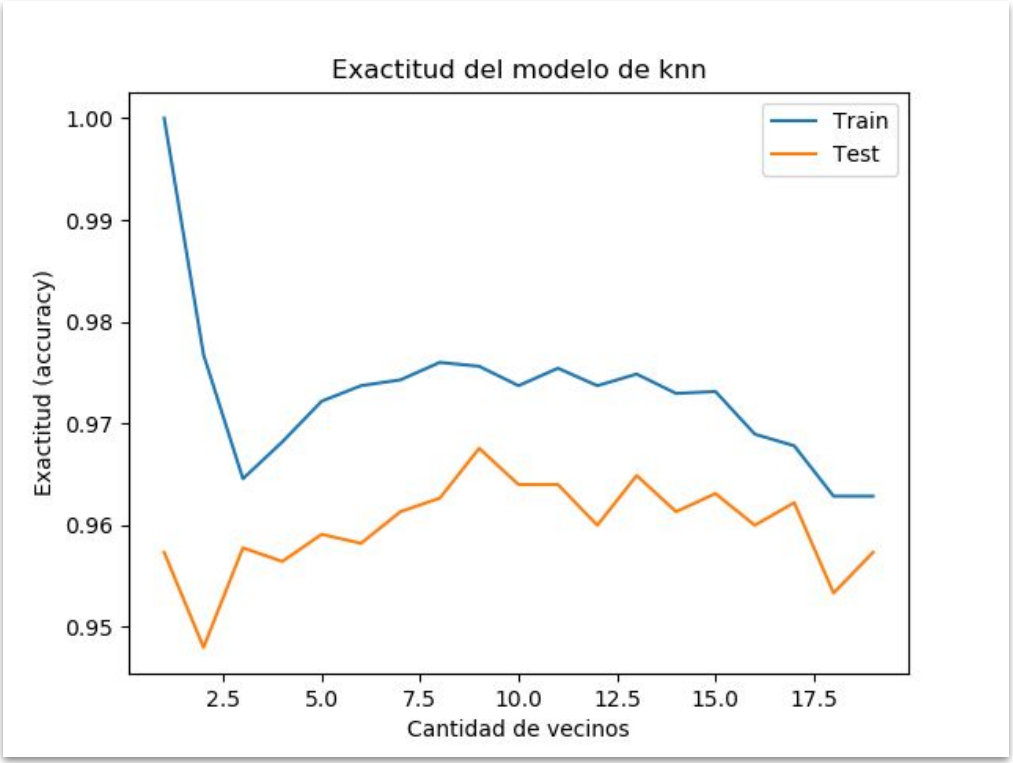
Entrenamos y evaluamos el modelo.

Ejemplos con Iris

¿Cuál es el mejor valor de k?

Evaluar los distintos valores de k y comparar

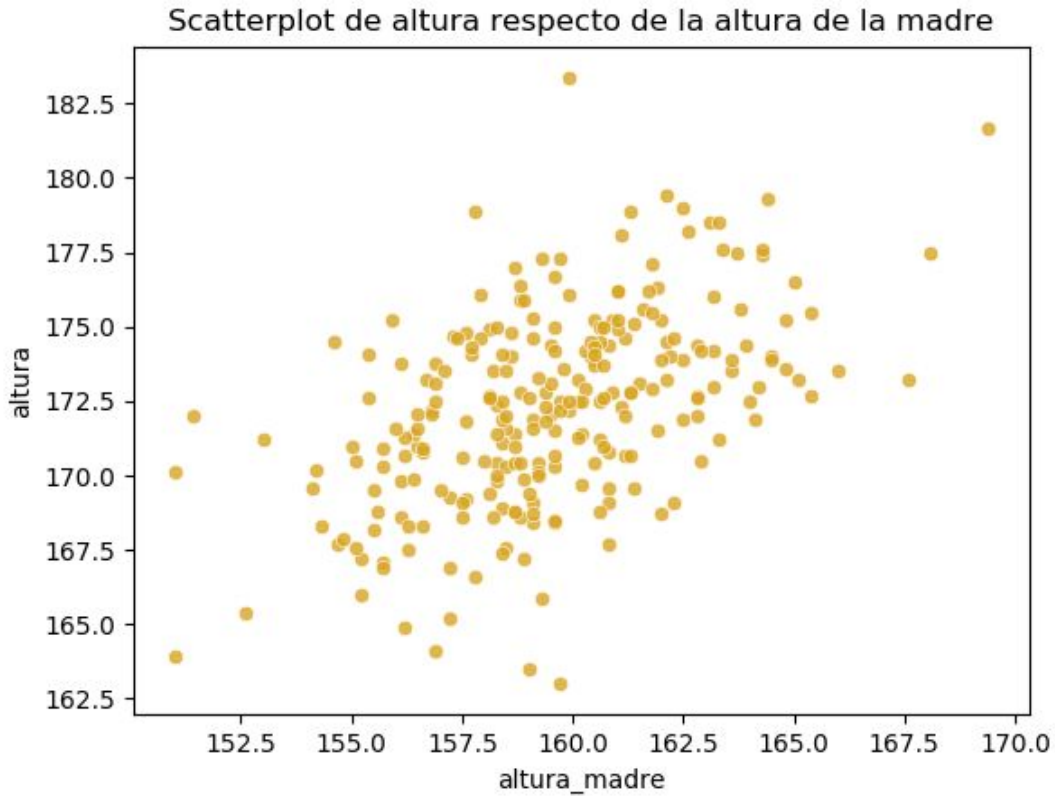
haciendo cross-validation.



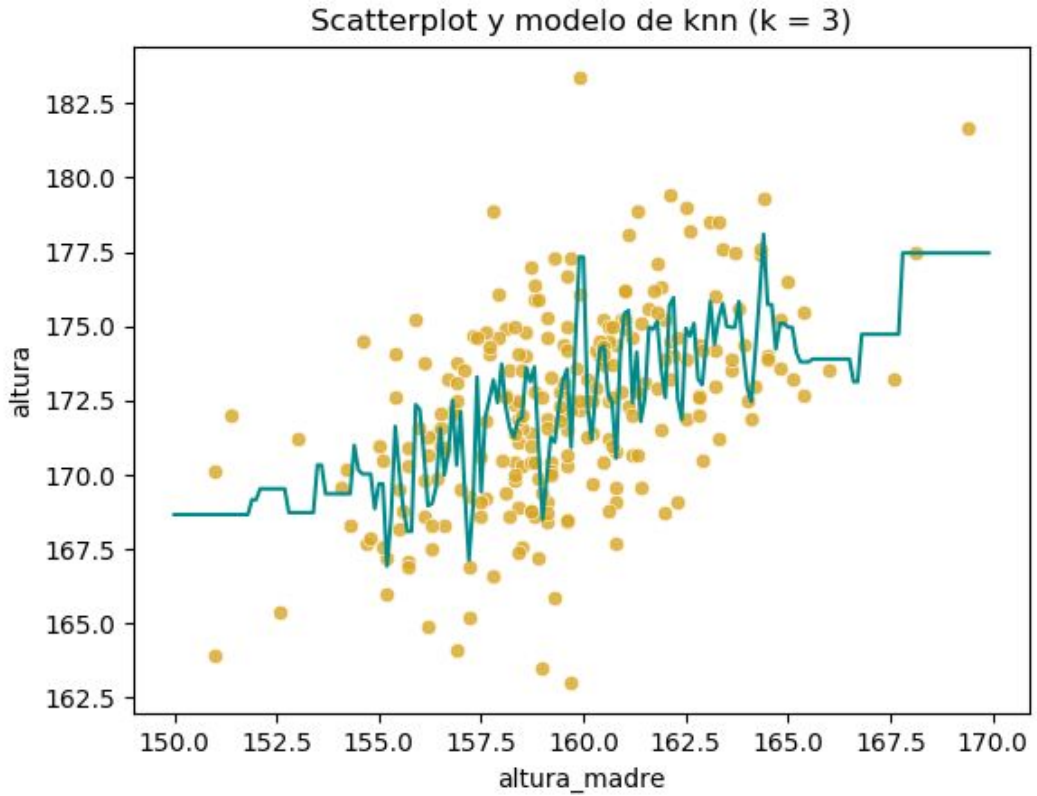
Ejercicio

+ Ajustar un modelo de kNN para predecir la altura que alcanzará el hijo, según los datos de la planilla.
+ Probar con distintos valores de k (por ejemplo 3, 5, 10), predecir la altura, si la madre mide 1.61.

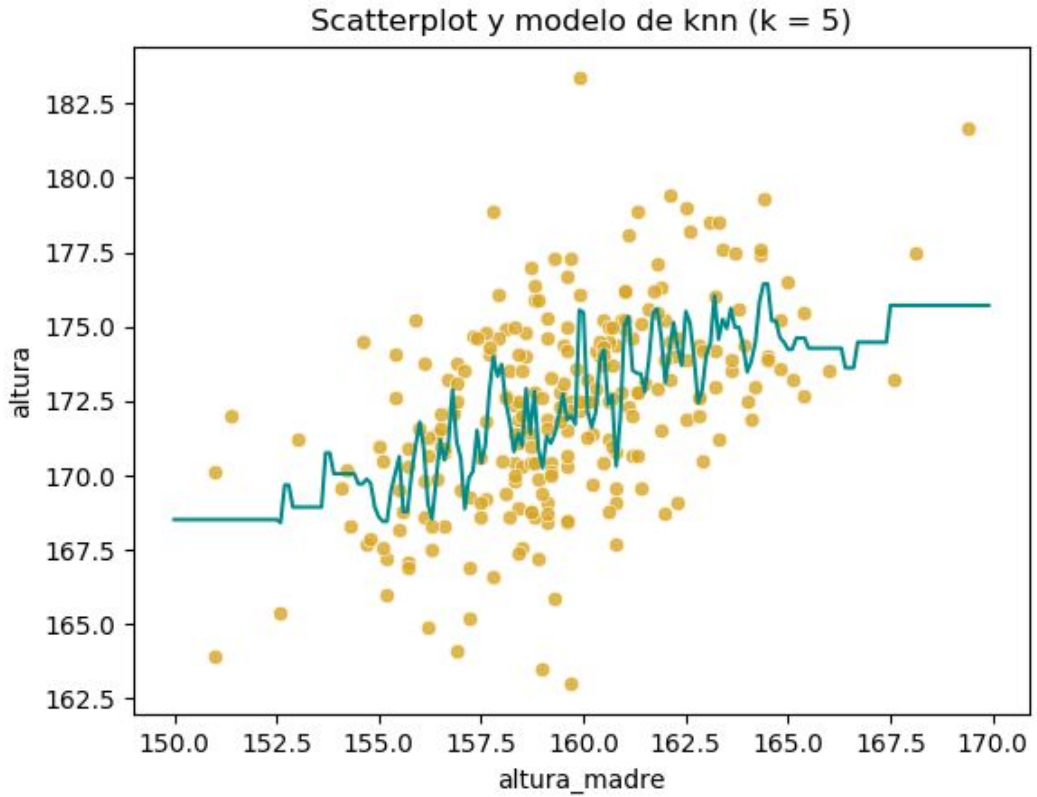
Ejemplo con datos simulados



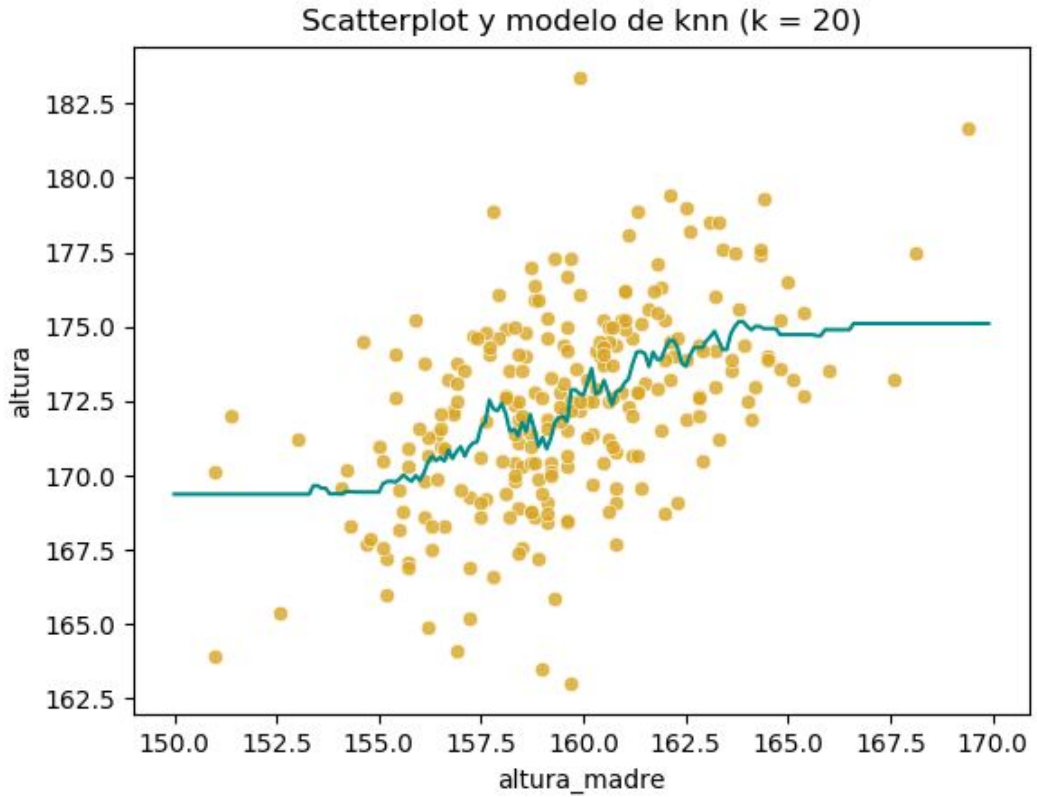
Ejemplo con datos simulados



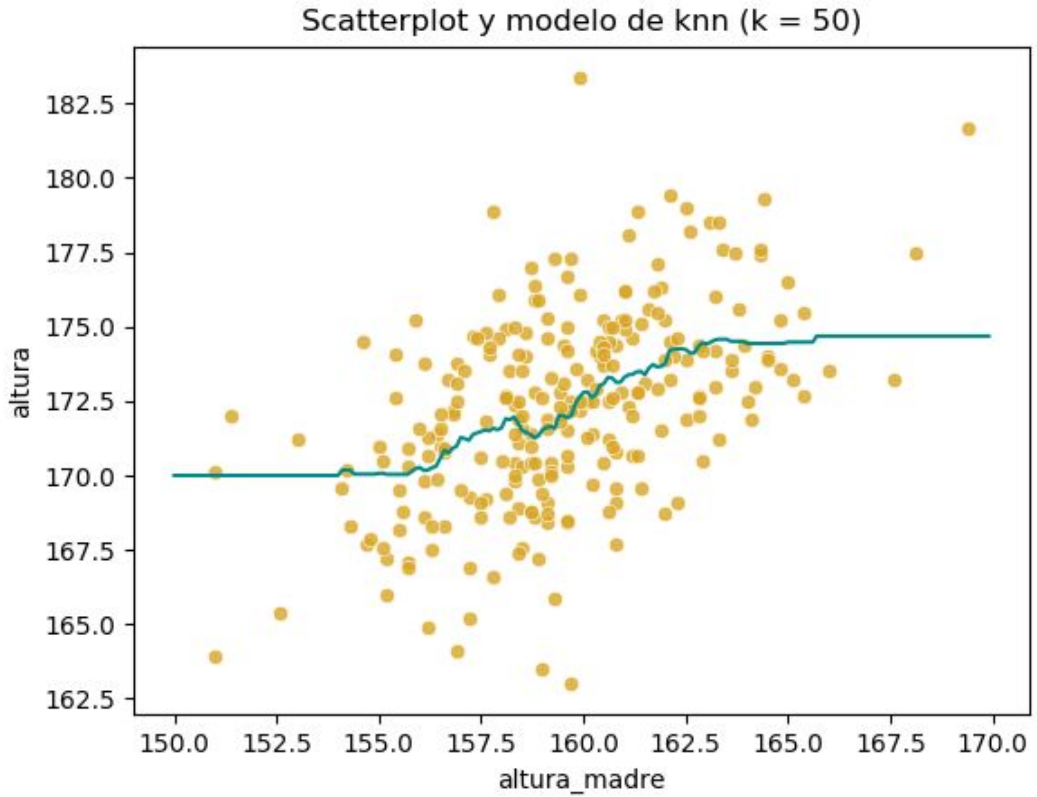
Ejemplo con datos simulados



Ejemplo con datos simulados



Ejemplo con datos simulados



Ejercicios

1. Pasamos al dataset de árboles. Cargar el csv.
2. Hacer un train-test split y hacer una clasificación con knn. Probar con distintos valores de k.
3. Cross-validation con k-folding: ajustar el modelo para cada valor de k dentro de un rango, y graficar la exactitud en función del k.
4. Reescalar los atributos para que tomen valores entre 0 y 1 y repetir. ¿Mejora la clasificación?

KNN involucra distancia
¿Cuál estamos usando?
¿Por qué puede impactar la escala?