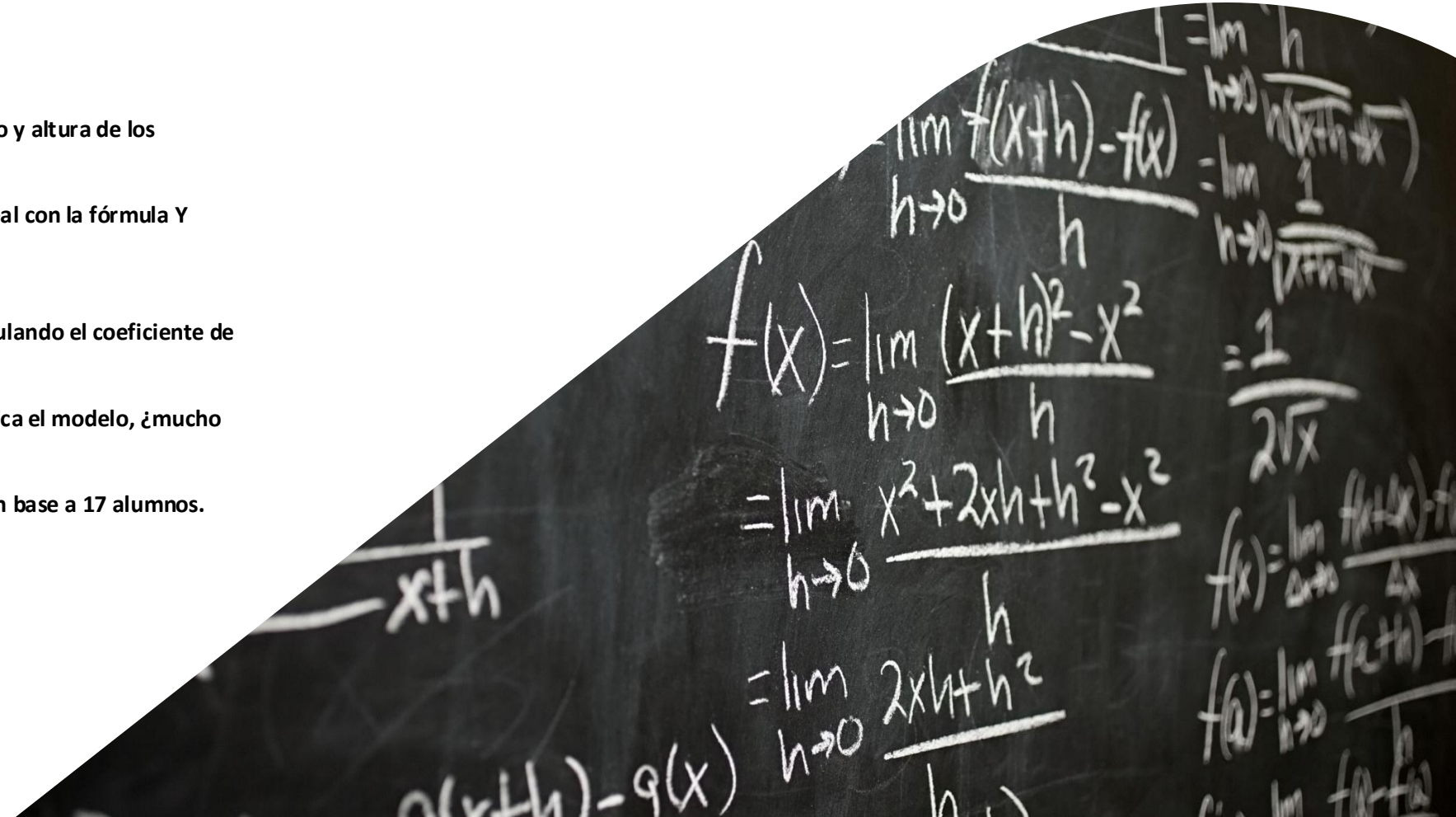
Estadística y Matemática

Andy Tsuen Kit Lui

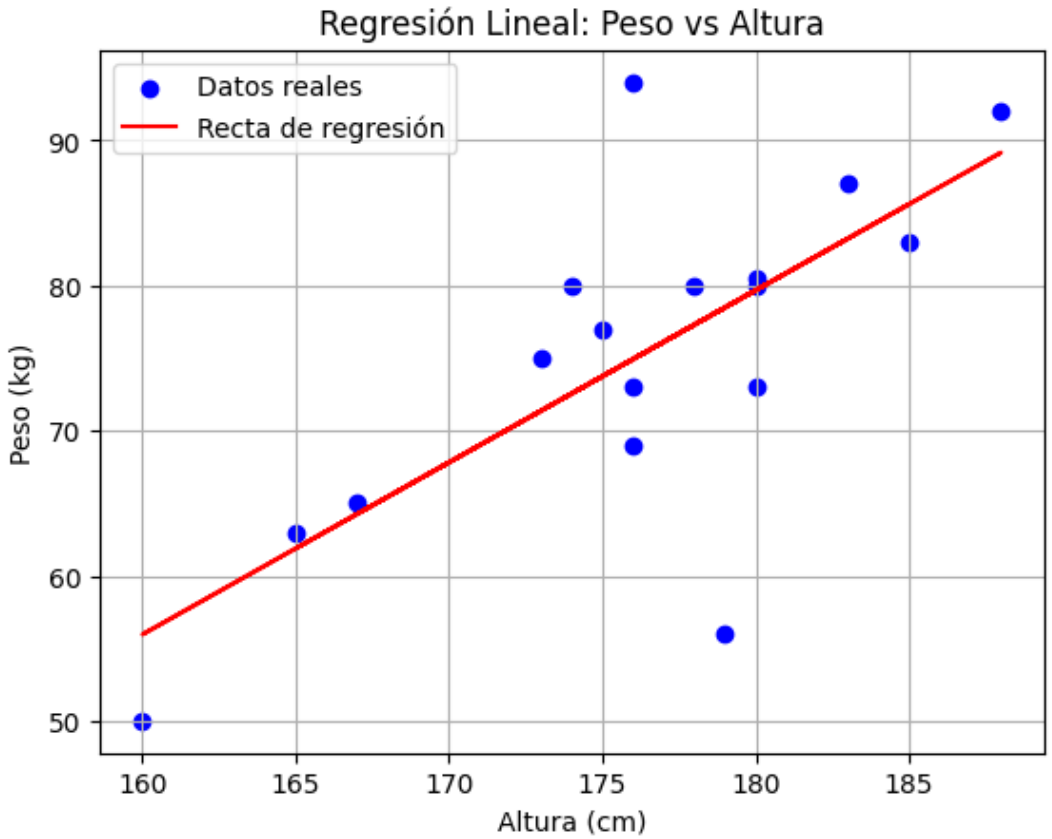
Modelo de regresión con estatura y peso

- Calcularemos la desviación típica del peso y altura de los alumnos.
- Ajustaremos un modelo de regresión lineal con la fórmula Y = a+bX+E
  donde X = altura & Y = peso
- Comprobaremos si se ajusta la recta calculando el coeficiente de determinación R^2.
- Explicaremos y evaluaremos cuánto explica el modelo, ¿mucho o poco?
- Generaremos una gráfica de regresión, en base a 17 alumnos.



Gráfica Modelo de Regresión Lineal

```
Desviación típica de la altura: 7.10 cm
Desviación típica del peso: 11.83 kg
Modelo de regresión: Y = -133.82 + 1.19X
Coeficiente de determinación R^2: 0.5063
```



[Enlace para el repositorio del código python]

Desviación típica

- La **desviación típica** (o desviación estándar) mide cuánto se dispersan los datos respecto a la media. La fórmula es:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- La **recta de regresión** es la línea que mejor ajusta los datos en un análisis de regresión lineal. Representa la relación entre dos variables, en este caso, **altura (X)** y **peso (Y)**.

Proceso Paso a Paso:

- **Media**: Se calcula el promedio de los datos.
- **Diferencias**: Se resta la media a cada valor.
- **Cuadrados**: Se elevan al cuadrado esas diferencias.
- **Promedio**: Se calcula la media de esos cuadrados.
- **Raíz cuadrada**: Se extrae la raíz cuadrada del resultado.

Coeficiente de determinación

- El **coeficiente de determinación R^2** indica qué tan bien se ajusta la **recta de regresión** a los datos.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

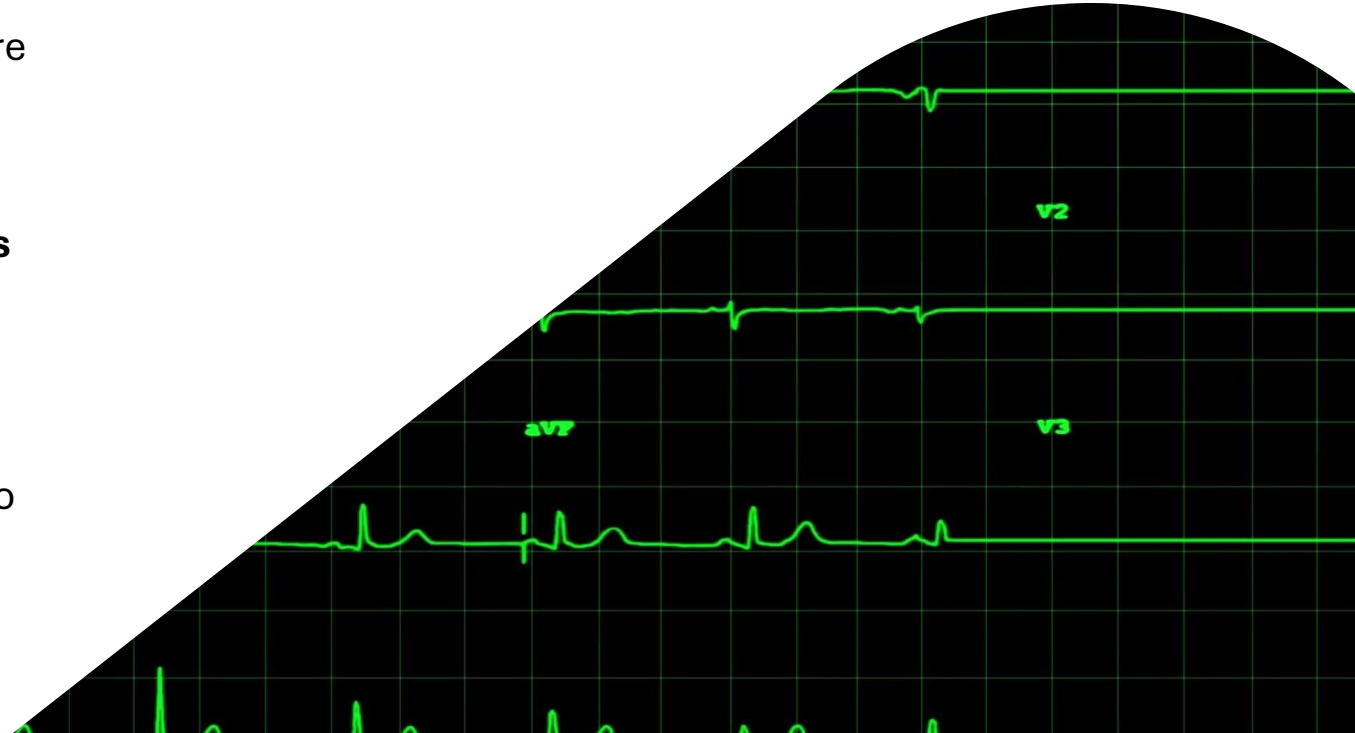
- **SSE (Suma de los errores al cuadrado):** mide el error entre los valores reales y los predichos.
- **SST (Suma total de cuadrados):** mide la variabilidad total de los datos respecto a la media.

- **$R^2$2 alto** (cercano a 1): el peso depende fuertemente de la altura.
- **$R^2$2 bajo** (cercano a 0): la relación entre peso y altura es débil.

- En este caso obtenemos 0.5063. El **50.63%** de la variabilidad del peso se explica por la altura.
- Por lo tanto, **no**, la recta de regresión **no se ajusta perfectamente** a los datos.

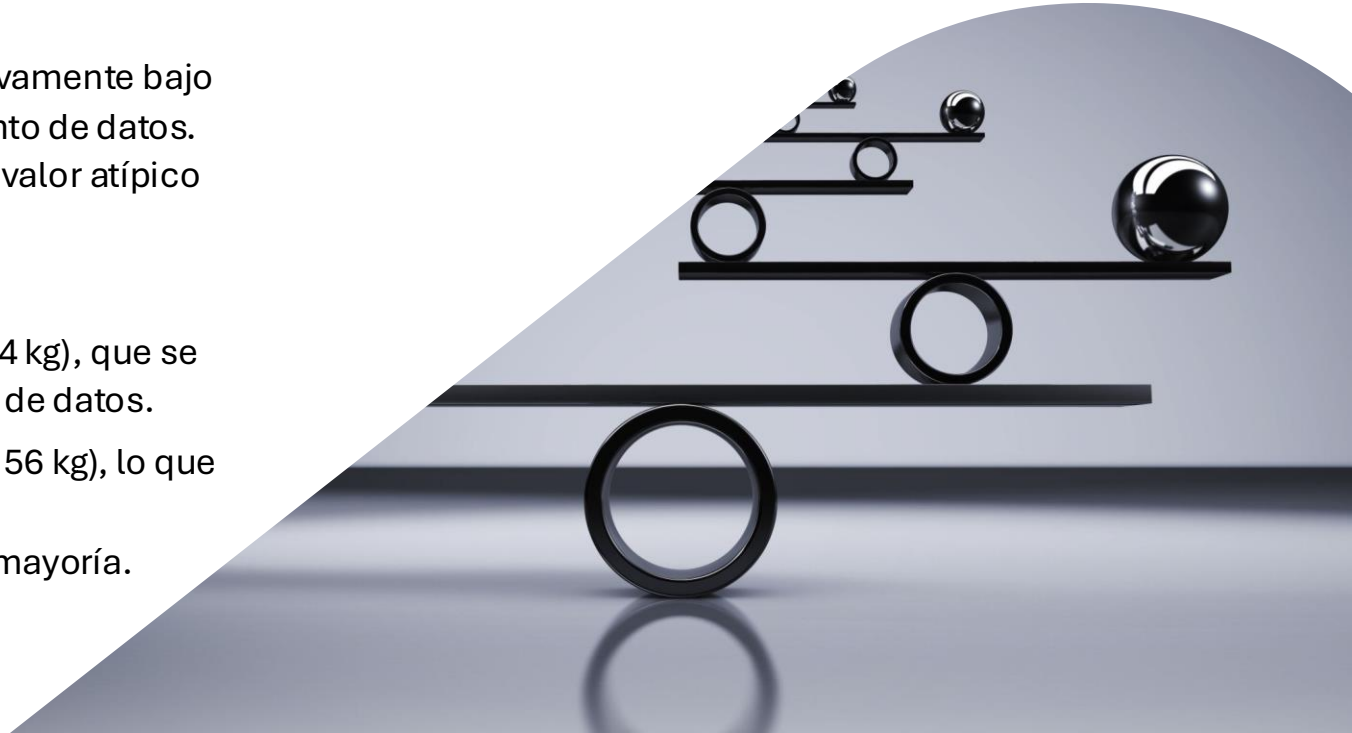
¿Cuánto explica el modelo, mucho o poco?

- La recta de regresión **no se ajusta perfectamente** a los datos.
- El coeficiente de determinación al ser 0.5063 sobre 1, está a la mitad aún para llegar a 1 y tener una fuerte relación entre peso y altura.
- Para mejorar el modelo sería ideal **incorporar más datos o variables** (ejemplo: edad, sexo, alimentación)
- Por lo tanto, **el modelo explica poco**, ya que la mitad de la variabilidad del peso no se explica solo con la altura.



Datos atípicos

- Los **valores atípicos** son aquellos que se encuentran lejos del comportamiento general del conjunto de datos.
- **Datos**:
  - **Alturas**: Los valores de altura oscilan entre 160 cm y 188 cm.
  - **Pesos**: Los pesos varían entre 50 kg y 94 kg.
- **Identificación de valores atípicos:**
  - **Altura**: Un valor de altura mínima de 160 cm es relativamente bajo comparado con la mayoría de las alturas en el conjunto de datos. Aunque no está muy alejado, podría considerarse un valor atípico porque está en el extremo inferior del rango.
  - **Pesos**:
    - Hay dos valores de peso superiores a 90 kg (92 y 94 kg), que se desvían del comportamiento general del conjunto de datos.
    - También hay dos valores por debajo de 60 kg (50 y 56 kg), lo que puede indicar que estas personas tienen un peso significativamente menor en comparación con la mayoría.



Coeficiente de correlación

- El **coeficiente de correlación** es una medida estadística que mide qué tan bien una variable se puede predecir a partir de la otra, siempre que la relación entre ellas sea lineal.
- Pasos:
  1. **Calcular las medias** de las dos variables
  2. **Calcular las diferencias** de cada valor con respecto a su media.
  3. **Multiplicar las diferencias** correspondientes para cada par de datos y luego sumarlas:
  4. **Calcular la sumatoria de los cuadrados de las diferencias** para cada variable:
  5. **Dividir la suma de los productos de las diferencias** entre el producto de las sumas de los cuadrados de las diferencias.
- Obtenemos 0.71, esto indica que existe una **correlación ligeramente positiva** entre ambas variables: a medida que la altura aumenta, el peso tiende a aumentar también, aunque no de manera perfectamente lineal.

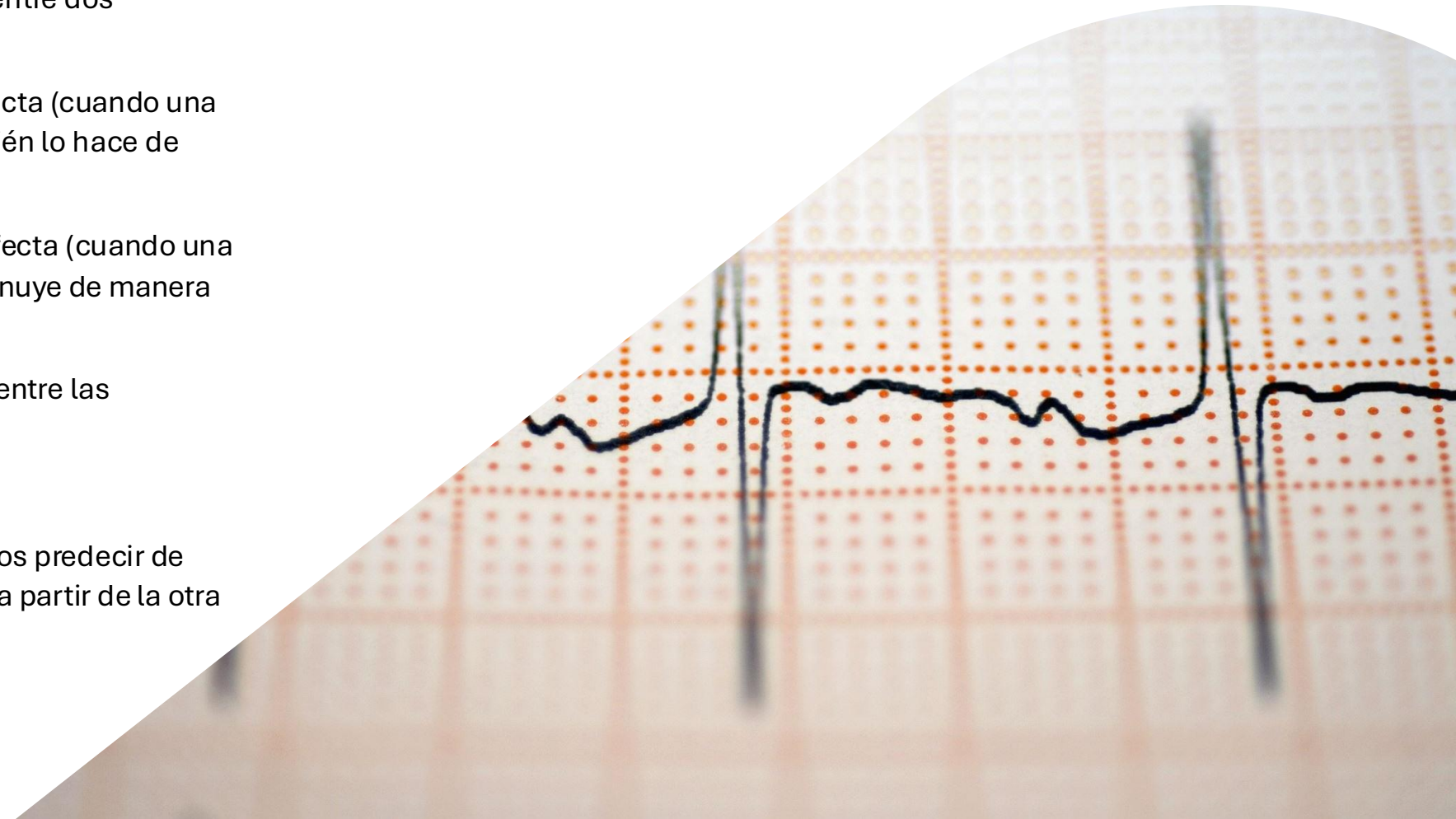
```
# Datos proporcionados (altura en cm, peso en kg)
alturas = np.array([180, 180, 160, 179, 165, 188, 176, 185, 178, 174, 180, 175, 167, 176, 183, 176, 173])
pesos = np.array([80, 73, 50, 56, 63, 92, 73, 83, 80, 80, 80.5, 77, 65, 69, 87, 94, 75])

# Calcular el coeficiente de correlación de Pearson
correlacion = np.corrcoef(alturas, pesos)[0, 1]
correlacion
```

¿Por qué cuando tengo coeficiente de correlaciones es R=0, sigo sin saber, sin tener información?

- El **coeficiente de correlación** mide la **fuerza y dirección** de la relación lineal entre dos variables. Va desde -1 hasta 1:
- **r = 1**: Correlación positiva perfecta (cuando una variable aumenta, la otra también lo hace de manera perfectamente lineal).
- **r = -1**: Correlación negativa perfecta (cuando una variable aumenta, la otra disminuye de manera perfectamente lineal).
- **r = 0**: No hay correlación lineal entre las variables.

- Por lo tanto, si **R = 0**, no podemos predecir de manera confiable una variable a partir de la otra utilizando una relación lineal.



Media VS Mediana

- La media y la mediana son dos medidas de tendencia central que se utilizan para describir un conjunto de datos.
- La **media** o promedio es el valor obtenido al sumar todos los valores de un conjunto de datos y luego dividir entre el número total de elementos.
- La **mediana** es el valor que ocupa la posición central en un conjunto de datos cuando están ordenados de menor a mayor (o de mayor a menor). Si el conjunto tiene un número impar de elementos, la mediana es el valor central. Si tiene un número par, la mediana es el promedio de los dos valores centrales.

Media VS Mediana
Ventajas e inconvenientes

Criterio	Media	Mediana
Definición	Promedio aritmético de los datos.	Valor central cuando los datos están ordenados.
Ventajas	Fácil de calcular.	No se ve afectada por valores atípicos.
	Utiliza toda la información.	Mejor para distribuciones sesgadas.
	Útil en distribuciones simétricas.	Representa el valor central en datos con extremos.
Inconvenientes	Sensible a valores atípicos (outliers).	No usa toda la información del conjunto.
	Puede no ser representativa en distribuciones sesgadas.	Puede ser más difícil de calcular en grandes conjuntos.

Ventajas e inconvenientes de la mediana con respecto a la media

LA **MEDIANA** PUEDE SER MEJOR QUE LA **MEDIA** EN CIERTOS CONTEXTOS DEBIDO A SU CAPACIDAD PARA:

- **Resiste la influencia de valores atípicos** (outliers)
- **Ideal para distribuciones sesgadas**
- **Funciona bien con datos que tienen un orden**, pero no son numéricos.
- **Representa el valor central**: Es el valor que divide el conjunto de datos ordenados en dos mitades iguales.
- **Menor sensibilidad a la variabilidad en los extremos.**

AUNQUE LA **MEDIANA** TIENE VENTAJAS SIGNIFICATIVAS EN CIERTAS SITUACIONES, TAMBIÉN PRESENTA VARIOS **INCONVENIENTES** RESPECTO A LA **MEDIA:**

- **No usa toda la información del conjunto**: No tiene en cuenta la magnitud de los valores, solo el valor central.
- **Es menos informativa** que la media en cuanto a la dispersión o la variabilidad de los datos.
- **No es útil para ciertos cálculos estadísticos** como varianza, desviación estándar o análisis de regresión.
- **Menos precisa en distribuciones simétricas**: En distribuciones equilibradas, la media es generalmente más útil.
- **Requiere ordenar los datos**, lo que puede ser más costoso en conjuntos grandes.
- **No refleja la variabilidad de los datos**, solo el valor central.

¿Cómo impacta en una recta de regresión el hecho de la presencia de outlier a la hora de estimar?

La presencia de **outliers** (valores muy alejados del resto de los datos) puede afectar de varias formas la línea de regresión:

- **Desvía la línea**: Los outliers pueden hacer que la línea de regresión se incline o se desplace, alejándose del patrón general de los datos.
- **Predicciones menos precisas**: Al influir en la posición de la línea, las predicciones que se hacen con el modelo pueden ser menos confiables y precisas.
- **Menor capacidad para explicar los datos**: La línea ajustada podría representar peor la relación entre las variables, haciendo que el modelo explique menos los datos.
- **Problemas con el análisis**: Los outliers pueden hacer que los errores del modelo sean más grandes o que se rompan algunas reglas importantes que debe cumplir la regresión.
- **Relación engañosa**: Un outlier puede hacer que parezca que hay una relación diferente entre las variables, como si una subiera cuando en realidad baja.

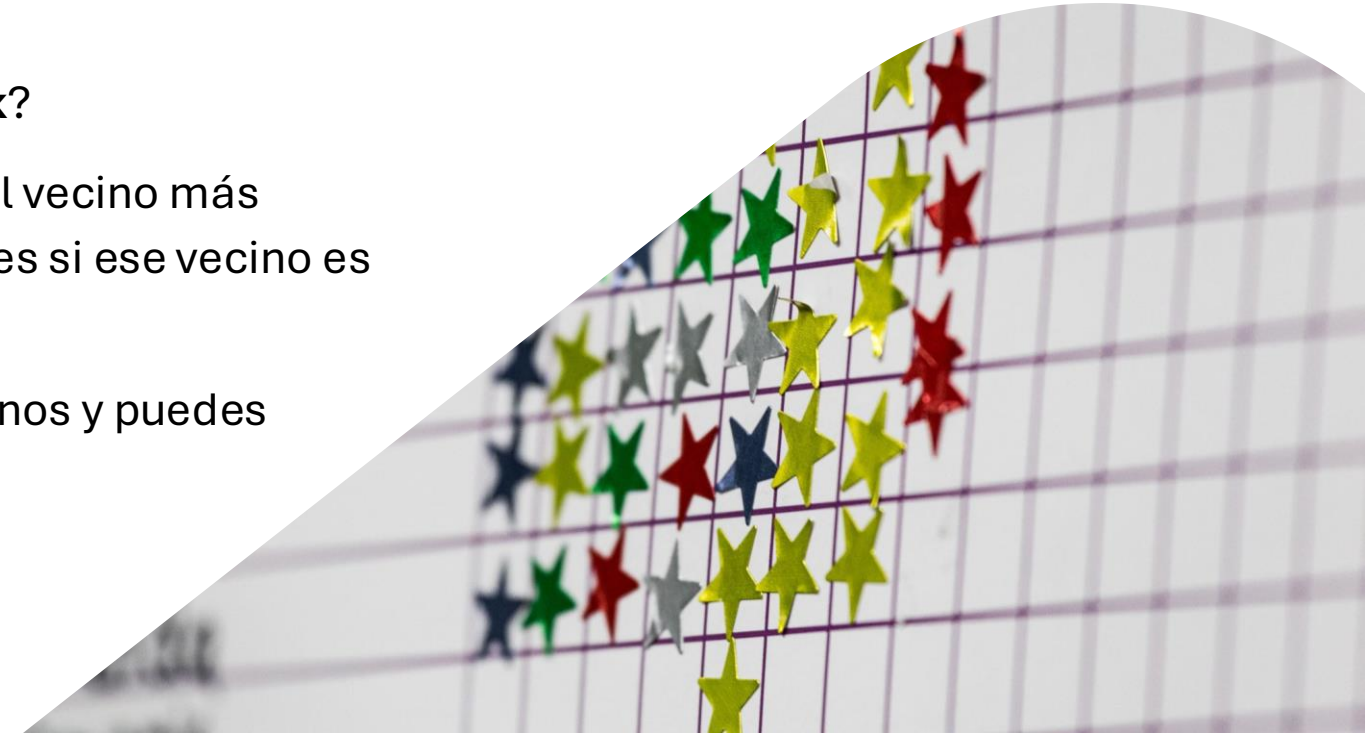
Para solucionar esto, se pueden detectar y revisar esos valores extraños, usar métodos más resistentes a los outliers o ajustar los datos para reducir su impacto.

Explicar con un ejemplo de aplicación del algoritmo de k vecino más próximo.

- Los **k vecinos** son los **k puntos más cercanos** a un nuevo dato que queremos clasificar o predecir. La letra **k** representa la cantidad de vecinos que vamos a considerar para tomar una decisión.

¿Por qué es importante elegir bien el valor de **k**?

- Si **k es muy pequeño** (como 1), solo miras al vecino más cercano. Esto puede hacer que te equivoques si ese vecino es un caso raro.
- Si **k es muy grande**, incluyes a muchos vecinos y puedes mezclar datos que no son tan relevantes.



Ejemplo 1. Amigos y fútbol.

- Imaginamos que tenemos un grupo de amigos y queremos saber si a una persona nueva le gusta el fútbol. No le vamos a preguntar directamente, sino que observas a las personas que están cerca de ella (sus vecinos). Si la mayoría de esos amigos cercanos juega al fútbol, probablemente a esa persona también le guste.

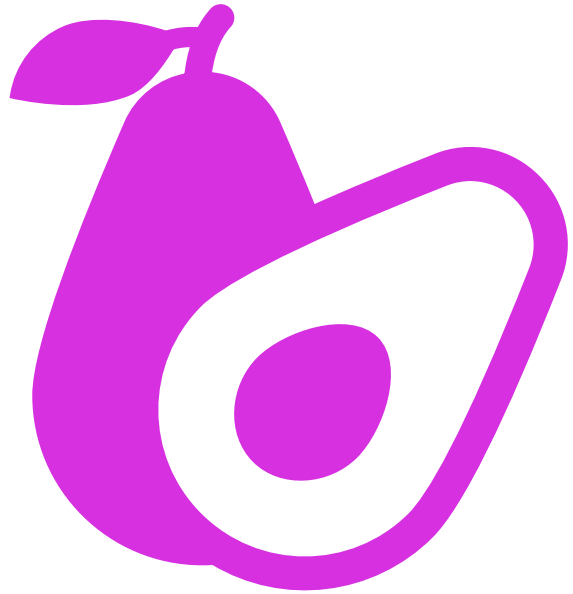
En el algoritmo **k-NN**, hacemos algo similar:

- **Calculamos la distancia** entre el nuevo dato y todos los datos que ya conocemos.
- **Elegimos los k más cercanos** (por ejemplo, los 3, 5 o 7 más próximos).
- **Observamos a qué grupo pertenecen** esos vecinos.
- **Decidimos por mayoría** a qué grupo pertenece el nuevo dato.

Ejemplo 2. Clasificar frutas

Imagina que tienes una cesta con manzanas y naranjas. Cada fruta tiene dos características:

- **Peso** (en gramos)
- **Color** (escala del 1 al 10, donde 1 es muy claro y 10 es muy oscuro)



Ya sabes qué frutas son manzanas y cuáles son naranjas basándote en estas dos características. Ahora, aparece una fruta nueva y quieres saber si es una manzana o una naranja.

¿Cómo funciona k-NN?

- **Dibuja los datos**: Imagina un plano donde el eje X es el peso y el eje Y es el color. Colocas puntos rojos para manzanas y puntos naranjas para naranjas.
- **Elige el valor de k**: Supongamos que elegimos **k = 3**. Esto significa que miraremos las 3 frutas más cercanas a la nueva fruta.
- **Busca los vecinos más cercanos**: Medimos la distancia (como si usaras una regla) entre la fruta nueva y todas las demás. Elegimos las 3 más cercanas.
- **Vota por la mayoría**: Si de esos 3 vecinos, 2 son manzanas y 1 es naranja, entonces clasificamos la fruta nueva como **manzana**.