



Facultad de  
Ingeniería

coursera

MIAD



Maestría  
en Inteligencia  
Analítica de Datos

La Universidad de los Andes como institución que almacena, y recolecta datos personales, requiere obtener su autorización para que de manera libre, previa, expresa, voluntaria y debidamente informada, permita a todas las dependencias académicas y/o administrativas, recolectar, recaudar, almacenar, usar, circular, suprimir, procesar, compilar, intercambiar, dar tratamiento, actualizar y disponer de los datos que han sido suministrados y que se han incorporado en distintas bases o bancos de datos, o en repositorios electrónicos de todo tipo con que cuenta la Universidad. Esta información es, y será utilizada en el desarrollo de las funciones propias de la Universidad en su condición de institución de educación superior, de forma directa o a través de terceros, a no ser que usted le manifieste lo contrario de manera directa, expresa, inequívoca y por escrito a la cuenta de correo electrónico dispuesta para tal efecto: [habeasdata@uniandes.edu.co](mailto:habeasdata@uniandes.edu.co) – Conozca más aquí: [www.uniandes.edu.co/datospersonales](http://www.uniandes.edu.co/datospersonales)

# MIAD



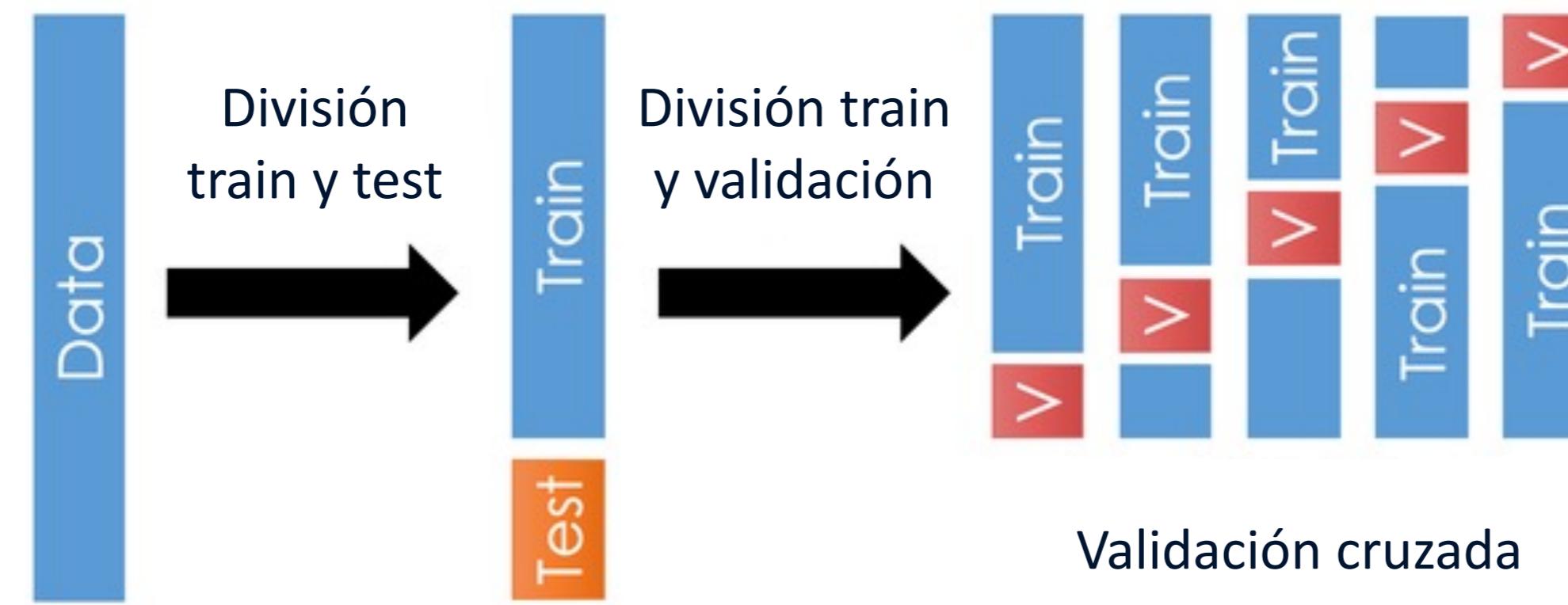
Maestría  
en Inteligencia  
Analítica de Datos

## Machine Learning y Procesamiento de Lenguaje Natural

Semana 2

# Dudas división Train y Test

Se debe separar sus datos en divisiones de entrenamiento y prueba para evitar que un modelo sobreajuste y para poder evaluar con precisión su modelo.



**Validación cruzada:** técnica para evaluar modelos mediante el entrenamiento en subconjuntos de los datos de entrenamiento y evaluarlos con el subconjunto complementario de los datos.

# Dudas división Train y Test

Se debe separar sus datos en divisiones de entrenamiento y prueba para evitar que un modelo sobreajuste y para poder evaluar con precisión su modelo.



**Validación cruzada:** técnica para evaluar modelos mediante el entrenamiento en subconjuntos de los datos de entrenamiento y evaluarlos con el subconjunto complementario de los datos.

# Matriz de confusión

		REAL	
		0	1
PREDICHO	0		
	1		

# Matriz de confusión

		REAL	
		0	1
PREDICHO	0	# Verdadero Negativos	
	1		# Verdaderos Positivos

# Matriz de confusión

		REAL	
		0	1
PREDICHO	0		# Falsos Negativos
	1	# Falsos Positivos	

# Matriz de confusión

		REAL	
		0	1
PREDICHO	0	# Verdadero Negativos	# Falsos Negativos
	1	# Falsos Positivos	# Verdaderos Positivos

# Accuracy

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

# Accuracy

$$\text{Accuracy} = \frac{\overbrace{VP + VN}^{\text{Total correctas}}}{VP + VN + FP + FN}$$

# Accuracy

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Total datos

# Problemas con el Accuracy

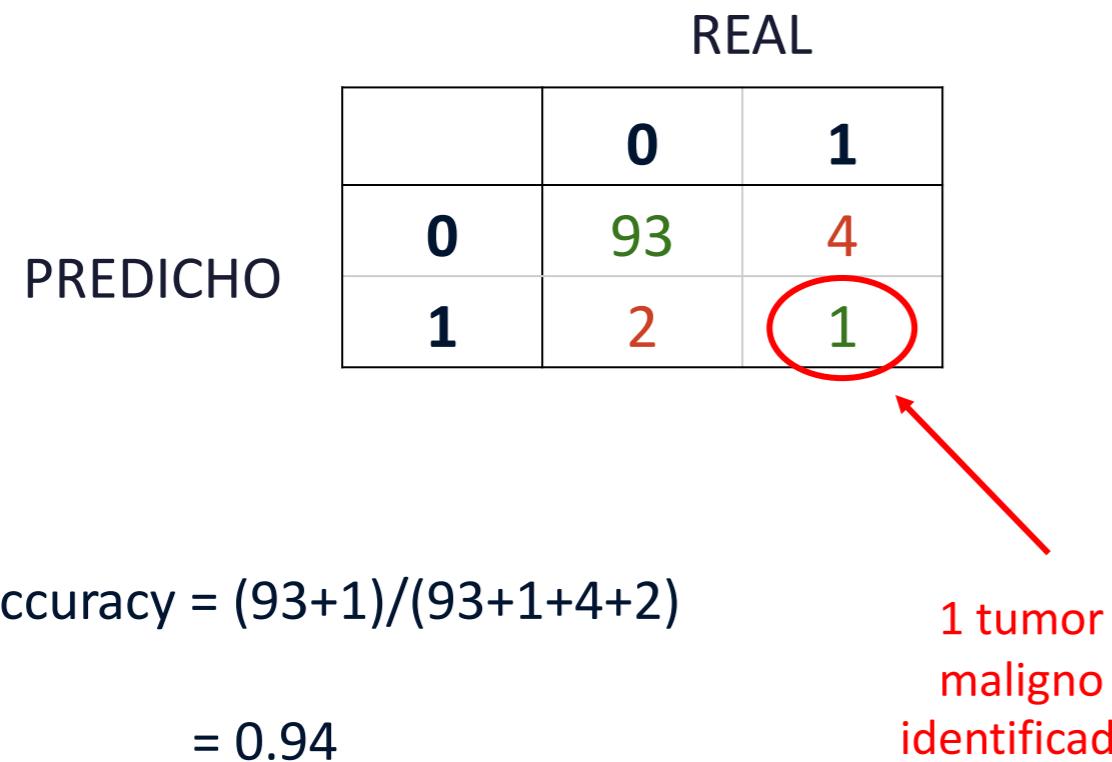
Problema: Identificar si un tumor es benigno o maligno

Tumor benigno: 0 – negativo  
95 casos

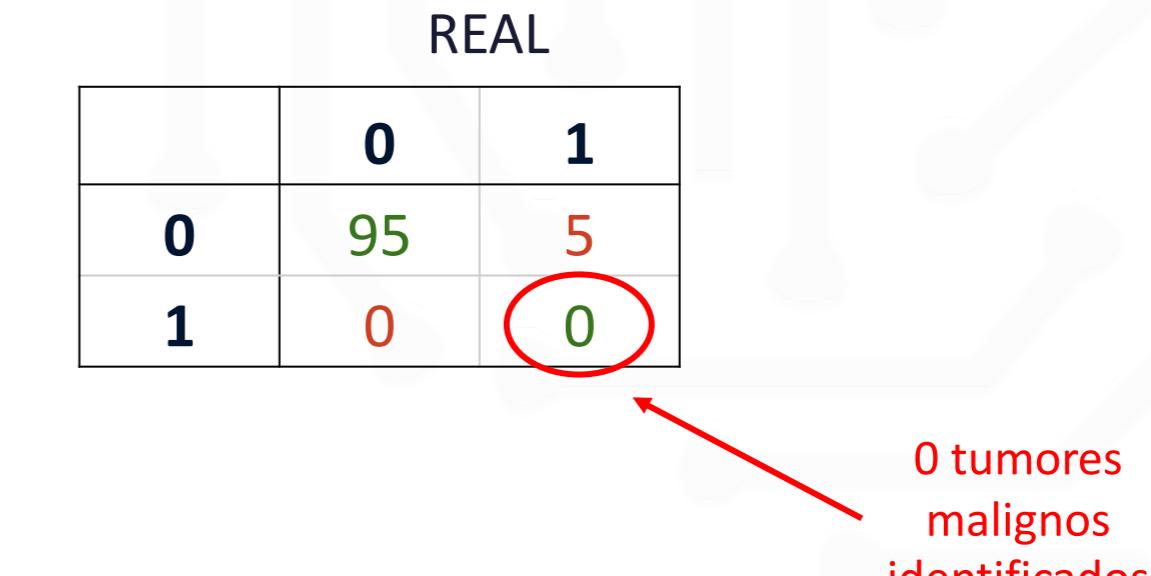
Tumor maligno: 1 – positivo  
5 casos

# Problemas con el Accuracy

MODELO 1:

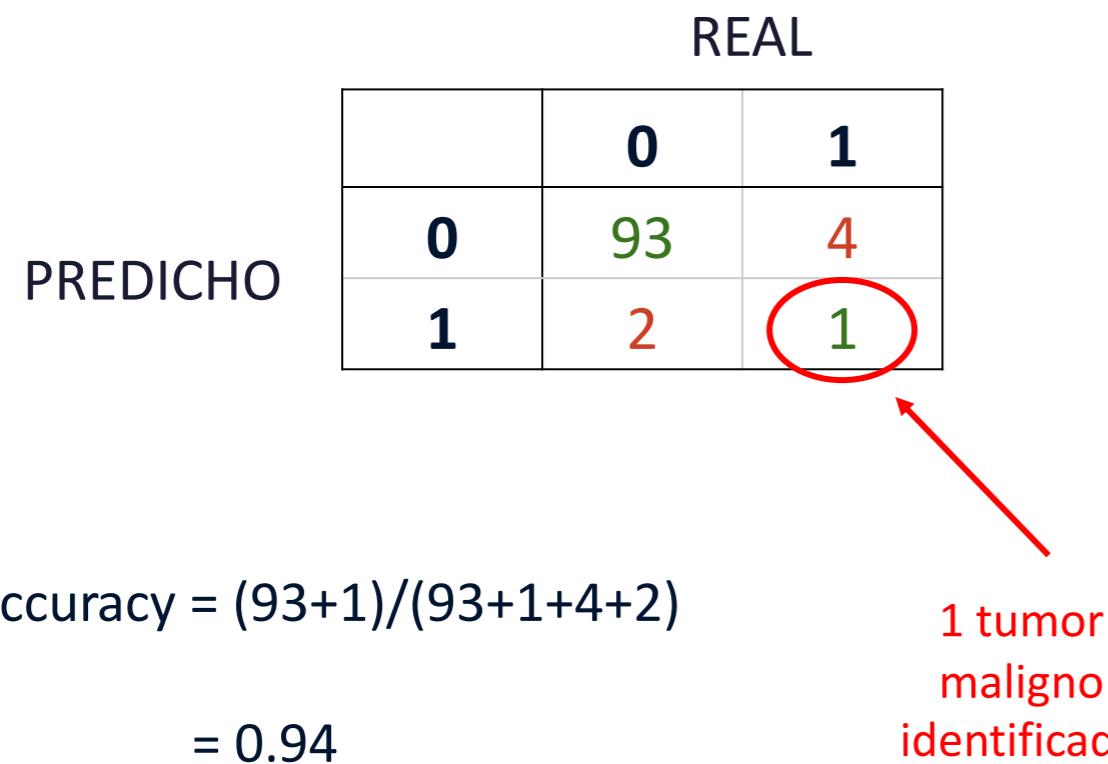


MODELO 2:

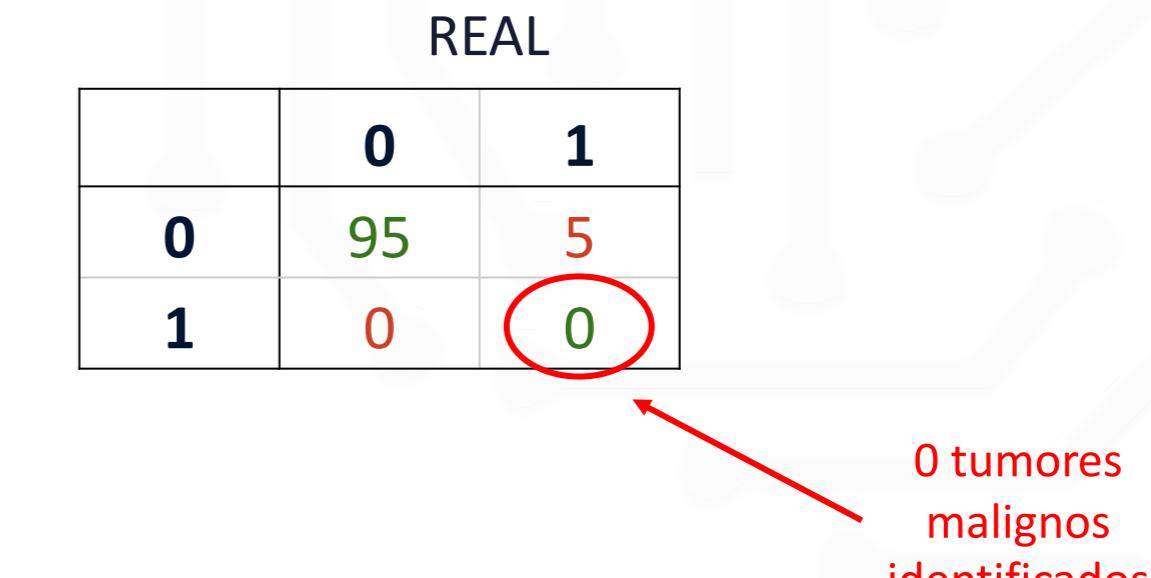


# Problemas con el Accuracy

MODELO 1:



MODELO 2:



El accuracy tiene problemas cuando el conjunto de datos está **desbalanceado**.

# Precisión

Identificados  
correctamente  
como positivos

$$\text{Precision} = \frac{VP}{VP + FP}$$

# Precisión

$$\text{Precision} = \frac{VP}{VP + FP}$$

Identificados  
como positivos  
por el modelo

# Precisión

$$\text{Precision} = \frac{VP}{VP + FP}$$

De todos los que el modelo clasificó como positivos, ¿qué proporción son realmente positivos?

## Precisión

Precisión es importante en casos como programas donde el gobierno va a entregar **100 VIVIENDAS** a familias necesitadas:

- Debemos identificar quién necesita vivienda.
- Son sólo 100 viviendas, entonces no le vamos a poder dar a todos los que necesiten.
- Pero, ¡no queremos darle a alguien que no la necesite!

# Recall

Identificados  
correctamente  
como positivos

$$\text{Recall} = \frac{VP}{VP + FN}$$

# Recall

$$\text{Recall} = \frac{VP}{VP + FN}$$

Total  
positivos

# Recall

$$\text{Recall} = \frac{VP}{VP + FN}$$

De todos los realmente positivos, ¿qué proporción clasificó el modelo como positivos?

# Recall

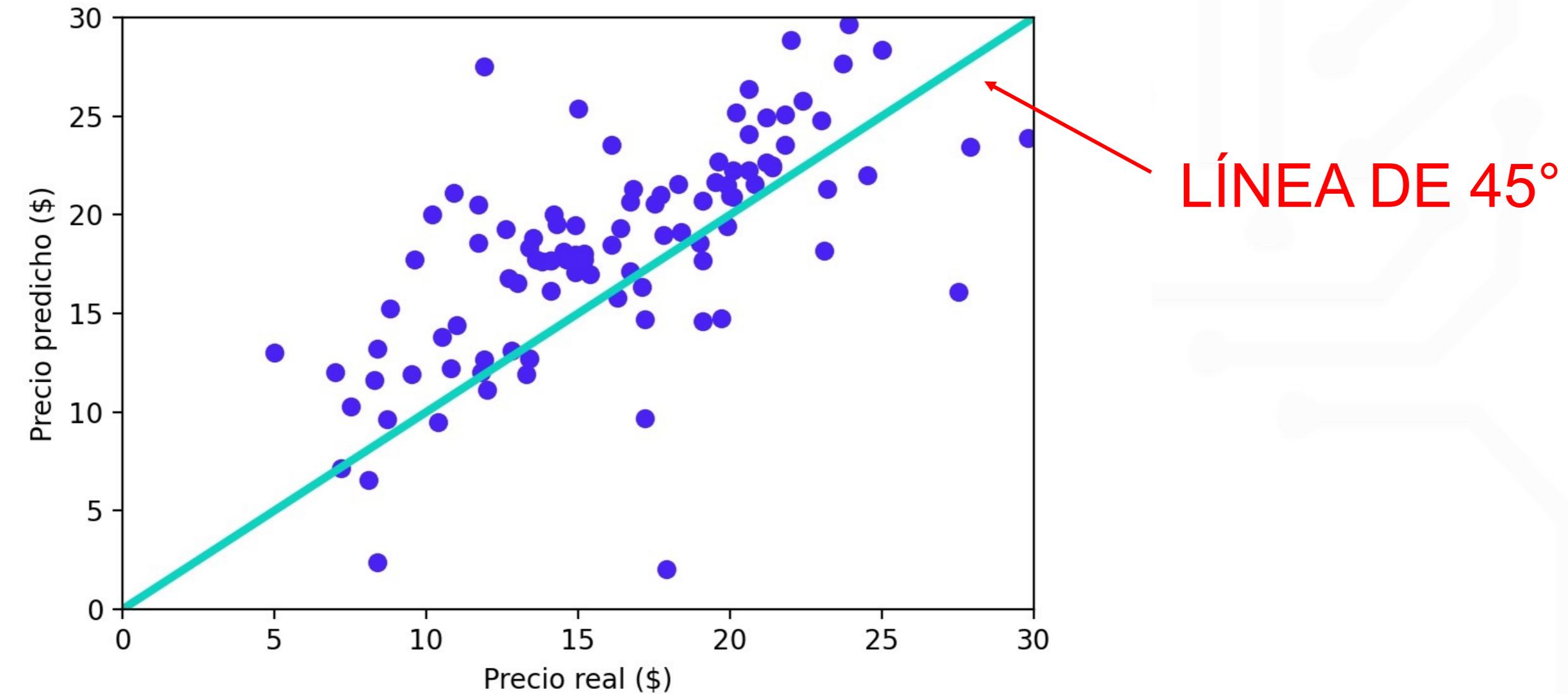
Recall es importante en casos como CORONAVIRUS:

- Debemos identificar quién necesita la vacuna.
- No importa que le pongamos la vacuna a alguien que no la necesite.
- Pero, ¡no ponérsela a alguien que sí la necesita puede ser muy costoso!

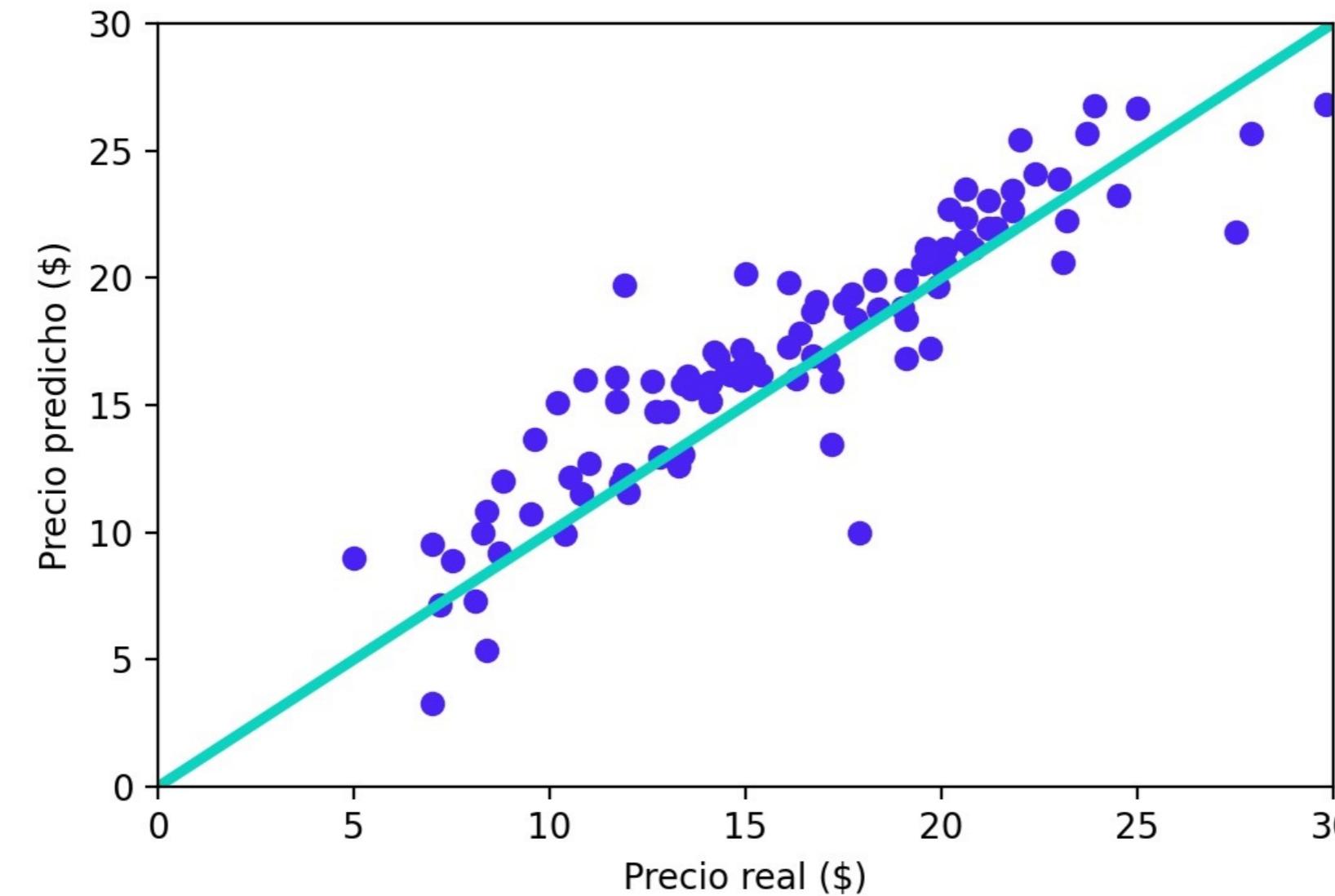
# F1-score

$$\text{F1-score} = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

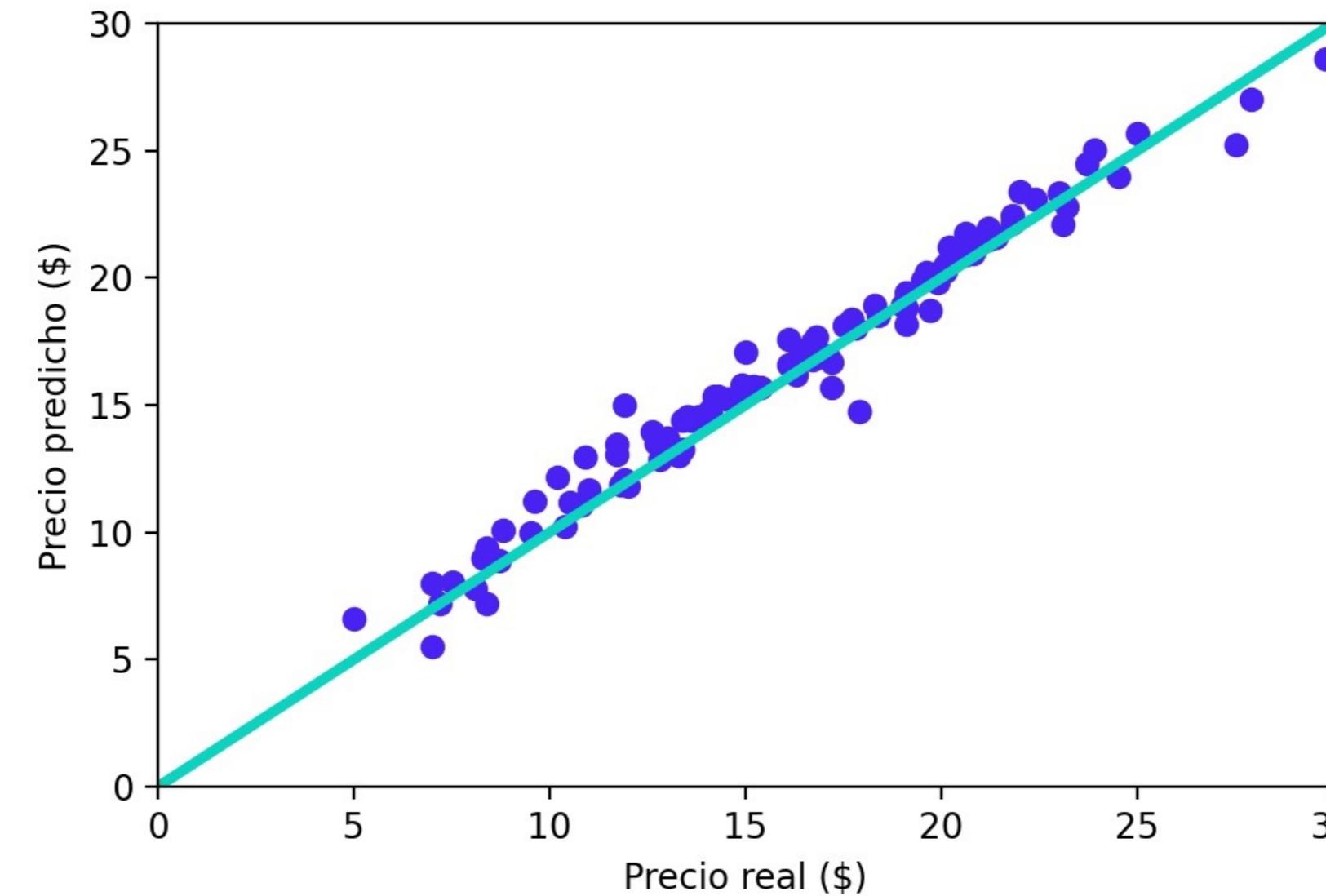
# Root Mean Squared Error



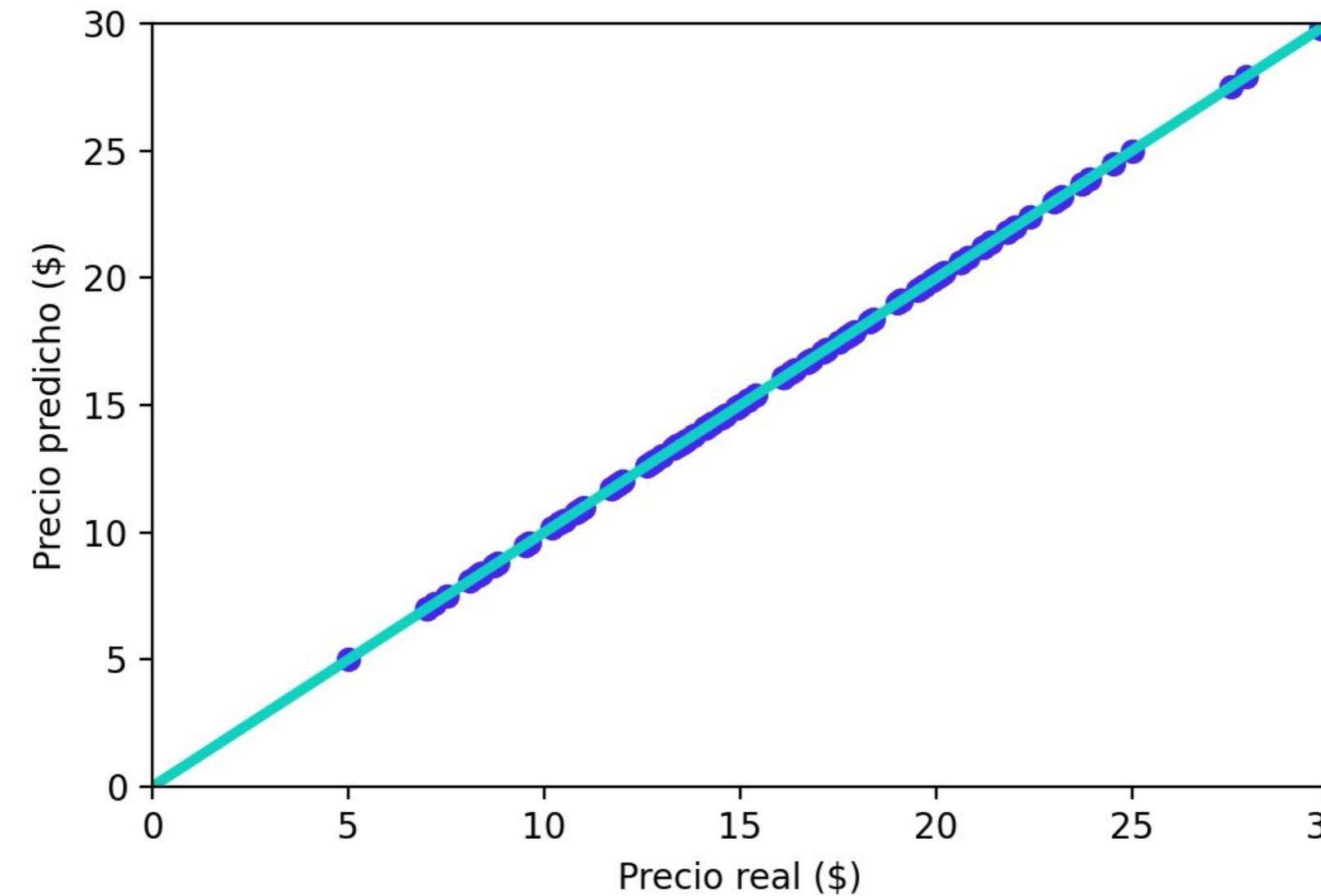
# Root Mean Squared Error



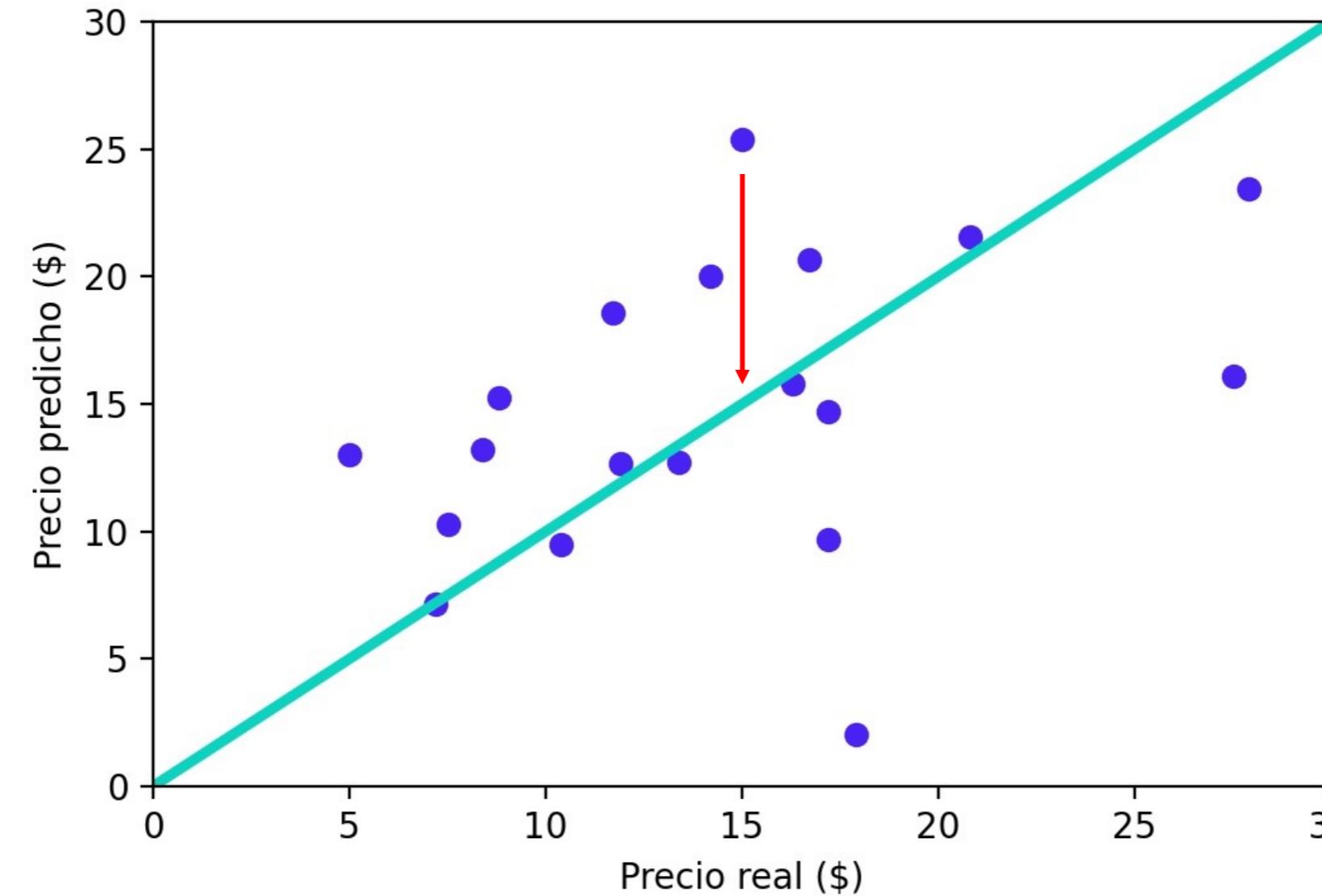
# Root Mean Squared Error



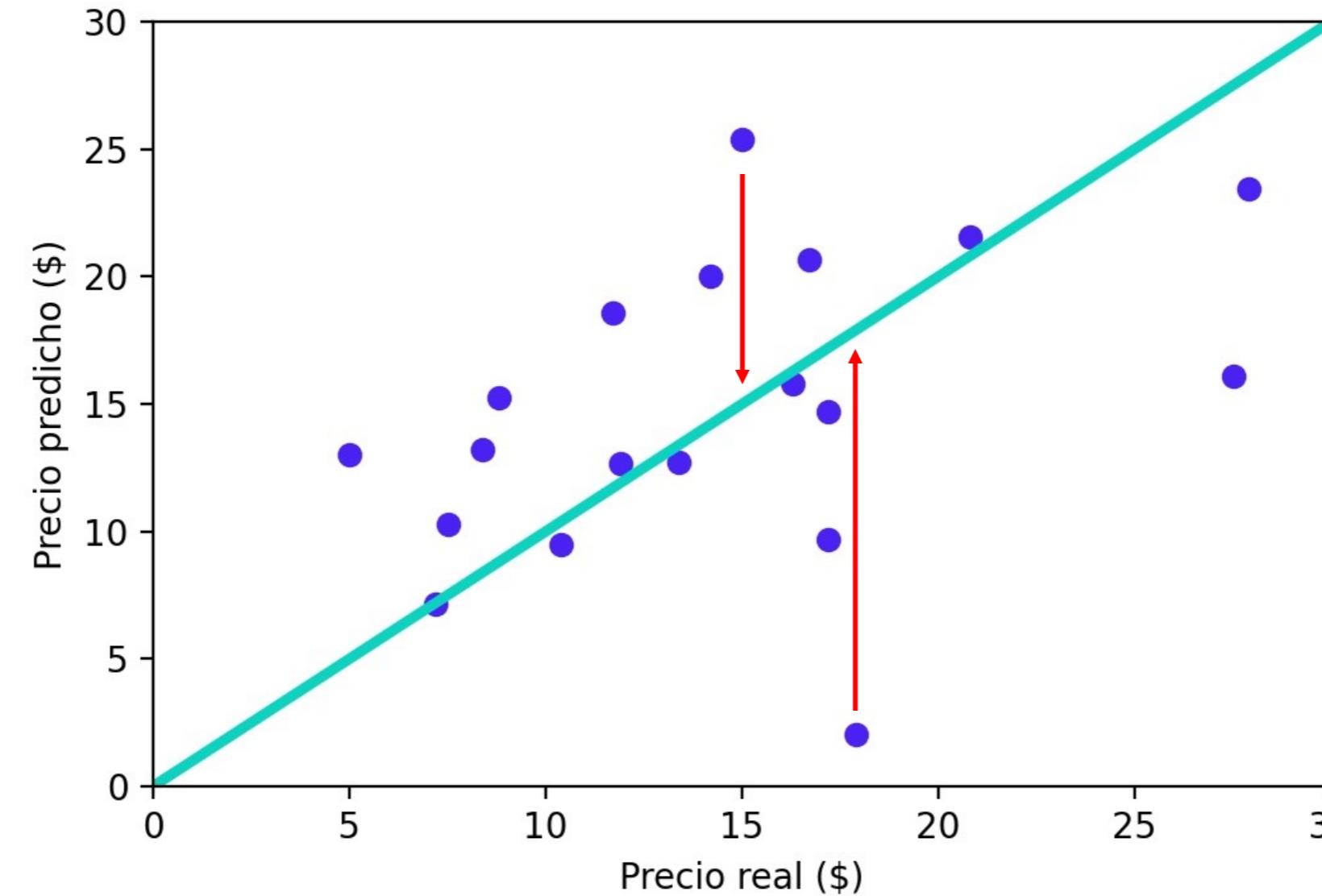
# Root Mean Squared Error



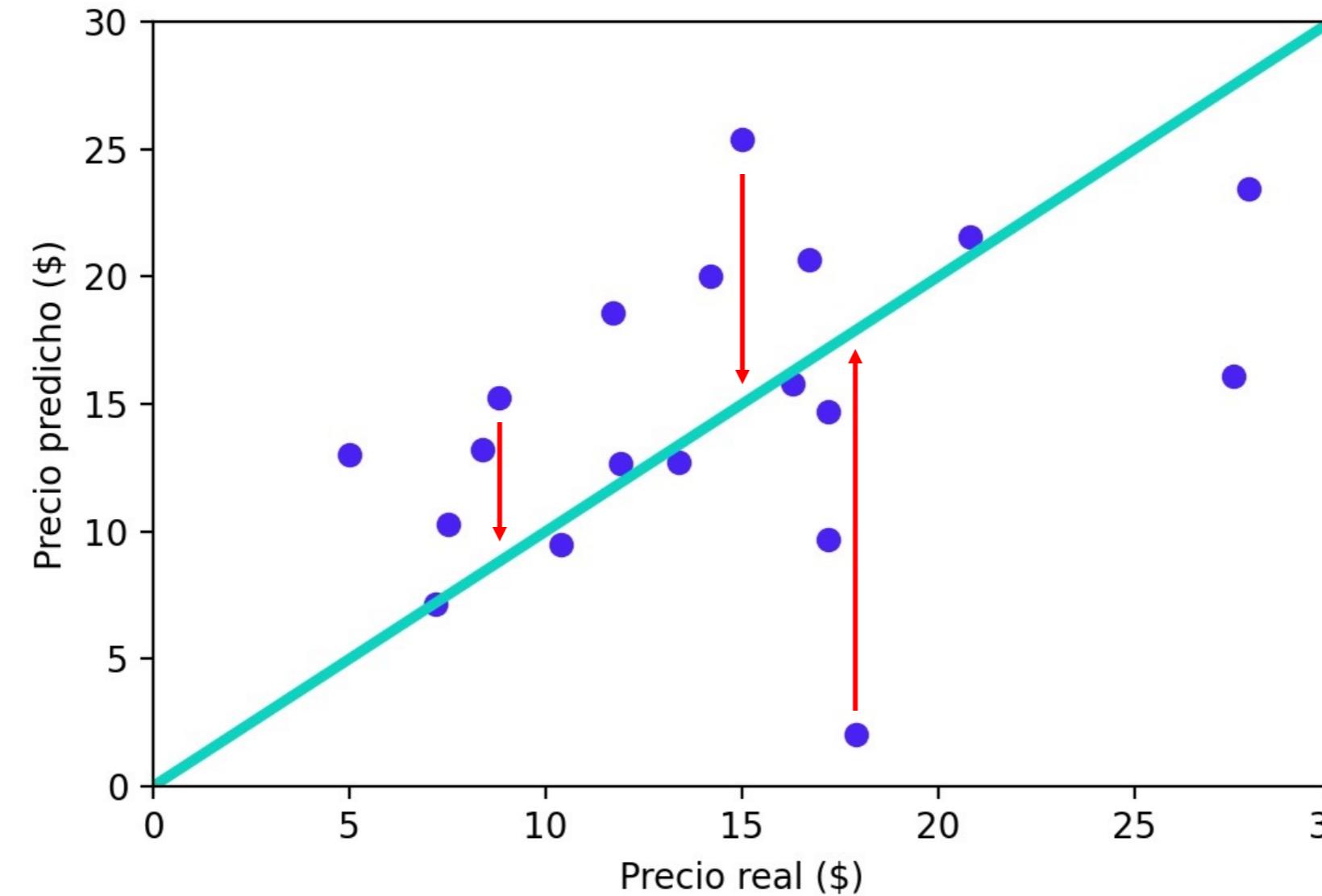
# Root Mean Squared Error



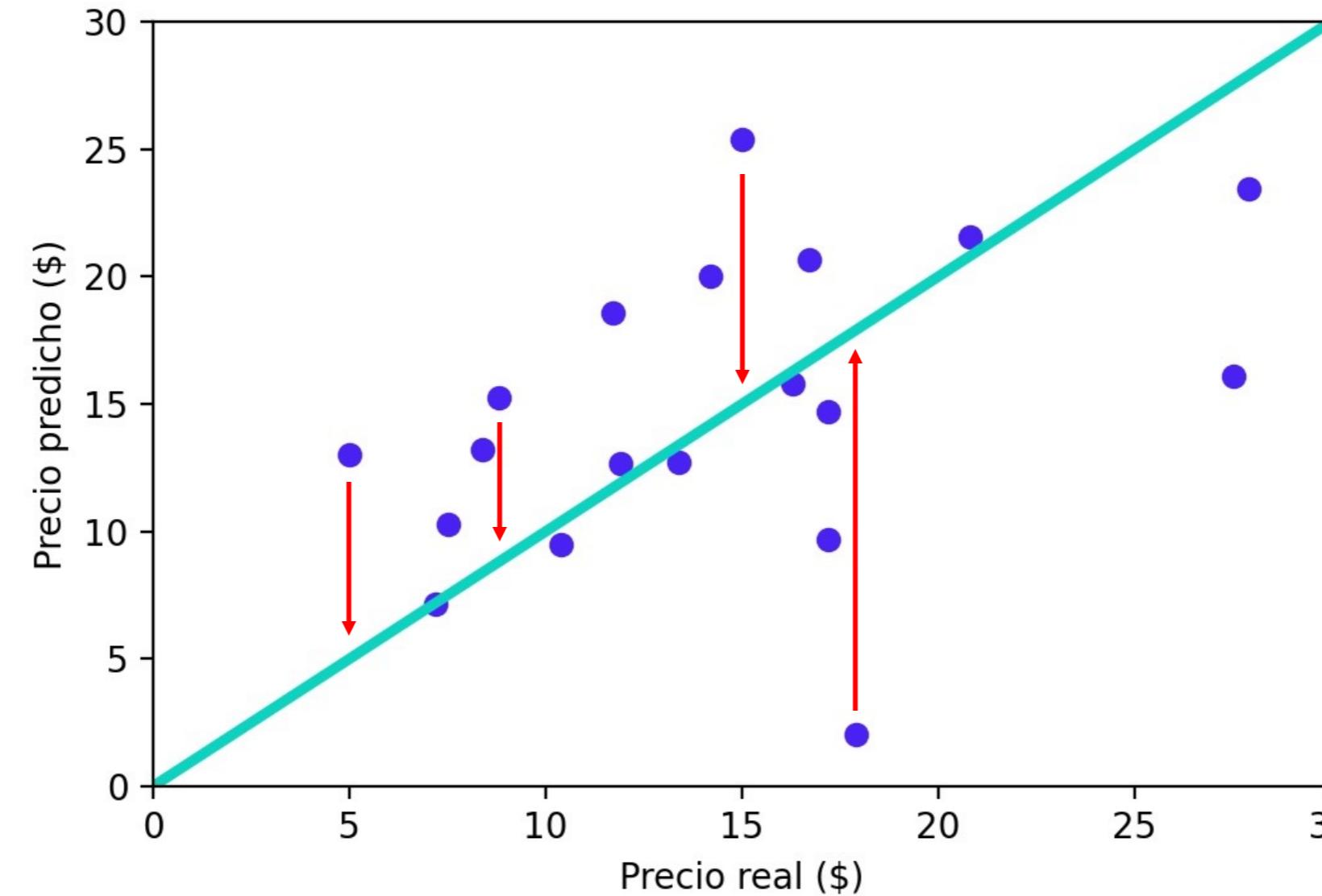
# Root Mean Squared Error



# Root Mean Squared Error



# Root Mean Squared Error



# Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

# Mean Absolute Error

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

# Mean Absolute Percentage Error

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{|y_i|}$$



# PREGUNTAS

# CONTÁCTANOS

PÁGINAS WEB - Ingeniería Industrial:

<https://industrial.uniandes.edu.co/es/programas-academicos/maestrias>

CORREO Maestría: [solicitudes-miad@uniandes.edu.co](mailto:solicitudes-miad@uniandes.edu.co)

## OTROS ENLACES DE IMPORTANCIA

Matrículas: <https://matriculas.uniandes.edu.co/>

Registro: <http://registro.uniandes.edu.co/>

Bloque Neón: <http://bloqueneon.uniandes.edu.co/>



AnalyticsUA



AnalyticsUA



AnalyticsUA



MIAD

