

M



Analítica avanzada





Agriculture



Variables y análisis





Tabla de análisis, variables y tipos de variables



ABT – Tabla base de Análisis

	Var1	Var2	Var3	...	X variables
1	321	15	Hombre	...	Audi
2	989	18	Mujer	...	Renault
3	726	21	Mujer	...	Chevrolet
4	654	9	Hombre	...	Seat
...	
n registros	198	43	Mujer	...	Renault

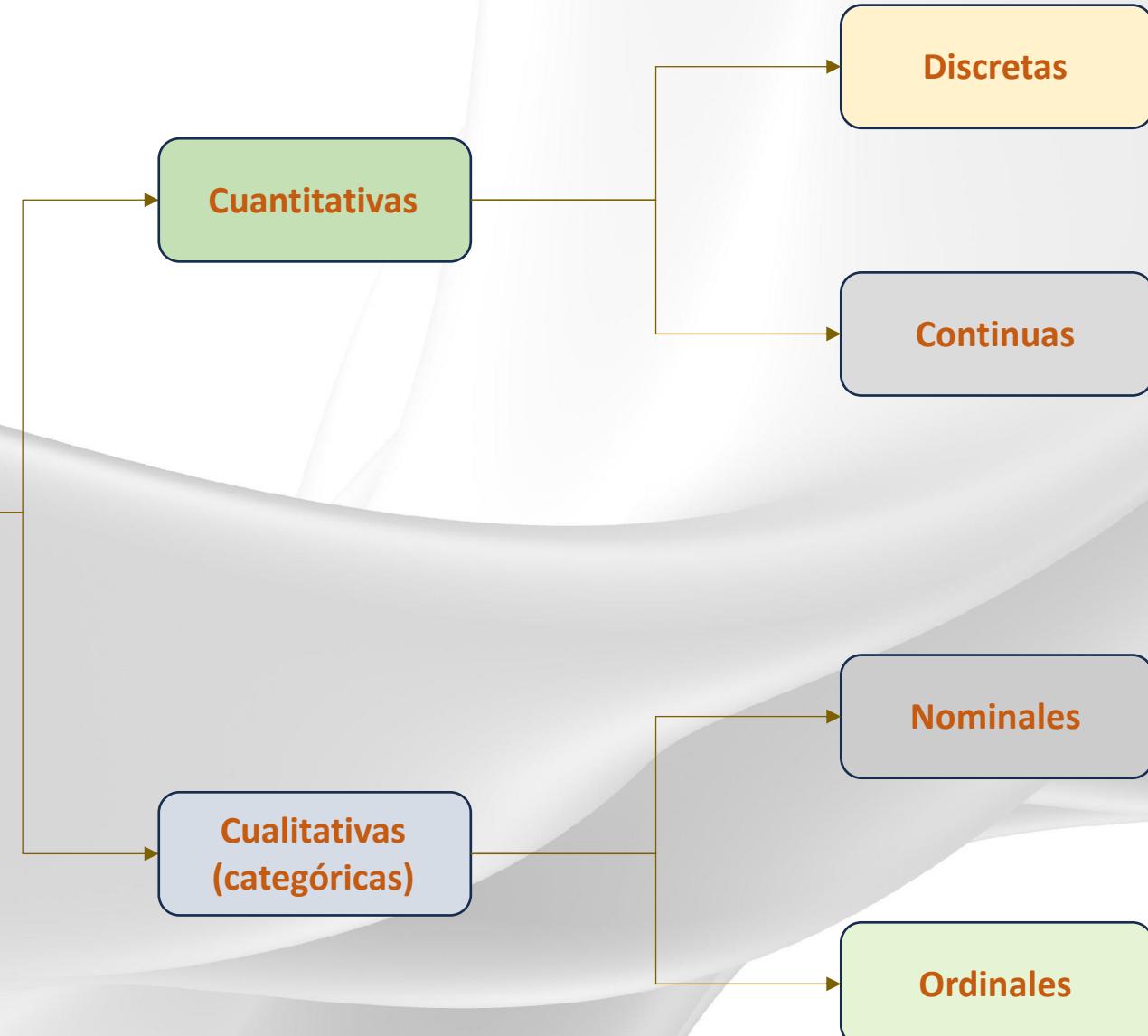
Atributos

- **Variables:** son las características o atributos que se miden.
 - Por ejemplo, si estás analizando clientes, las variables podrían ser su edad, género, ingresos, etc.
- **Registros:** Son las entradas de datos.
 - Por ejemplo, cada cliente sería un registro.



Tipos de variables

Variables



Variables cuantitativas

Representan magnitudes

¿Cómo saber si una variable es cuantitativa?

Prueba de la diferencia:
Restar dos valores de esta variable

¿Tiene sentido la diferencia?

- ✓ Edad: 60 años – 25 años
- ✗ Número de ID: 1019467 - 667893

Variables cuantitativas

Discretas

Dentro de un rango. Solo pueden tomar ciertos **valores determinados**

Número de hijos de la persona.

Cantidad de materias perdidas.

Continuas

Dentro de un rango, **sin restricción a valores determinados**. Pueden ser infinitos.

Edad del personal de Telefónica.

Estatura de las personas que trabajan en Morato.

¿Cómo saber si una variable es continua?

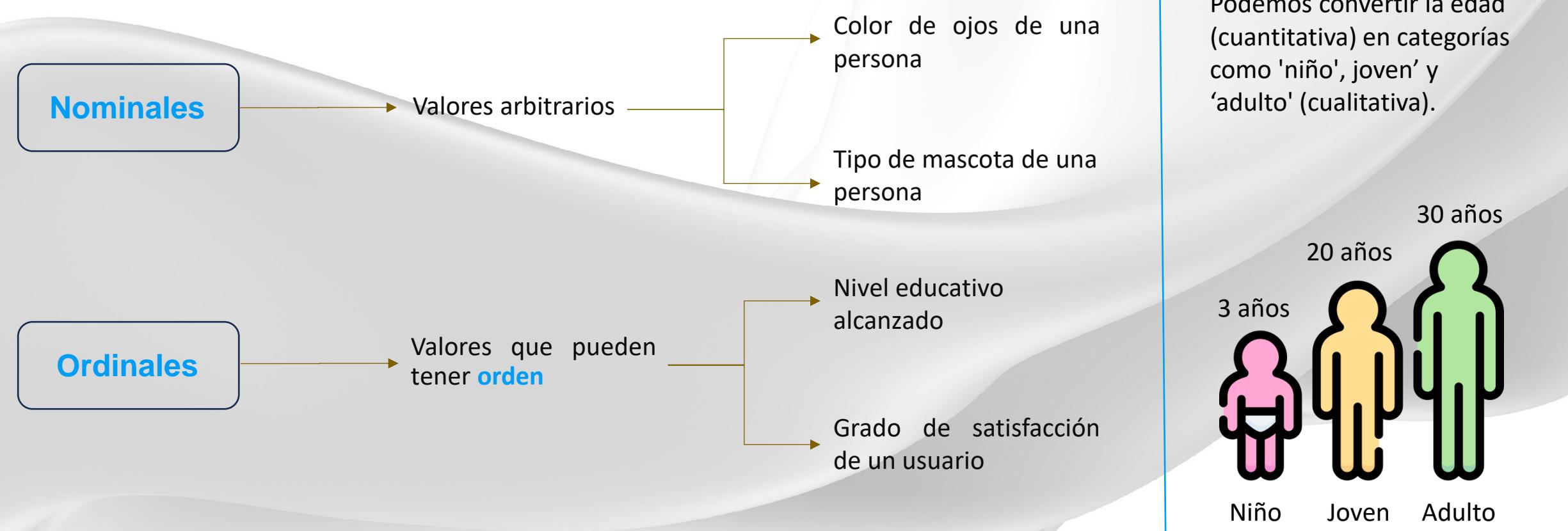
Si tiene sentido encontrar un valor intermedio, como 1,6 metros de altura, entonces es una variable **continua**.



✓ Continua: Calorías quemadas por día

✗ Discreta: Cantidad de bicicletas

Variables cualitativas





Análisis univariado

Análisis univariado

¿En qué consiste?

El análisis univariado con medidas estadísticas examina una sola variable a la vez para entender su comportamiento.



Variables cualitativas

Color del Carro	Frecuencia
Amarillo	4
Gris	3
Rojo	2
Azul	1

Análisis univariado

Variables cuantitativas



Imagina que tienes la altura de 10 personas...

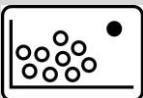
Promedio



Si sumas todas las alturas y las divides entre 10, obtienes la media.

Esto te dice cuál es el 'promedio' de altura de esas personas.

Datos atípicos



Si alguna de estas personas 2 metros, esta será catalogada como un valor atípico por ser muy diferente a la población.

Mediana

Si ordenas las alturas de mayor a menor y eliges el de la mitad, esa será la mediana.



Desviación estándar



Permite ver la diferencia promedio de altura en nuestra población.



Análisis bivariado

Análisis bivariado

¿En qué consiste?

Medidas que permiten determinar relaciones entre dos variables.

Variables cualitativas

Ejemplo: tabla cruzada entre género y rango etario

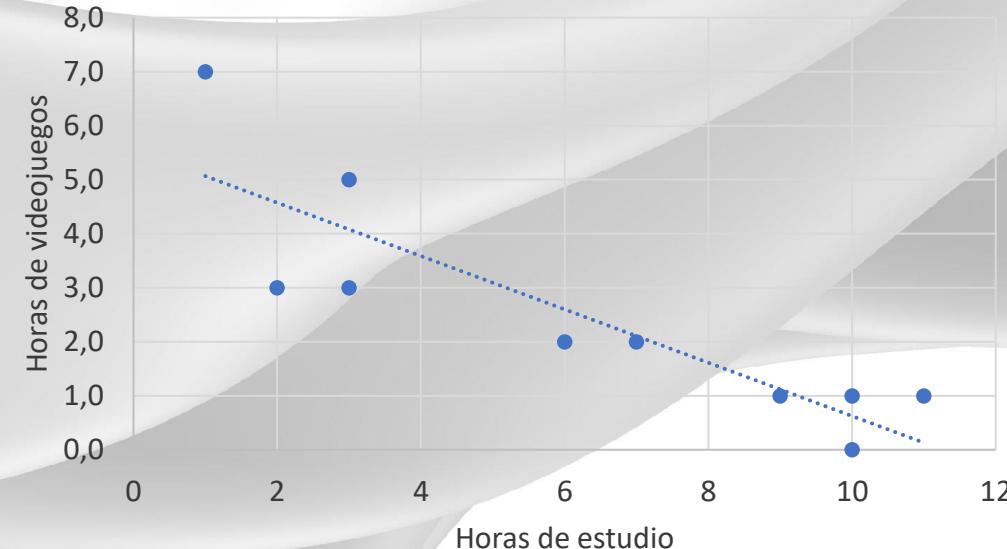
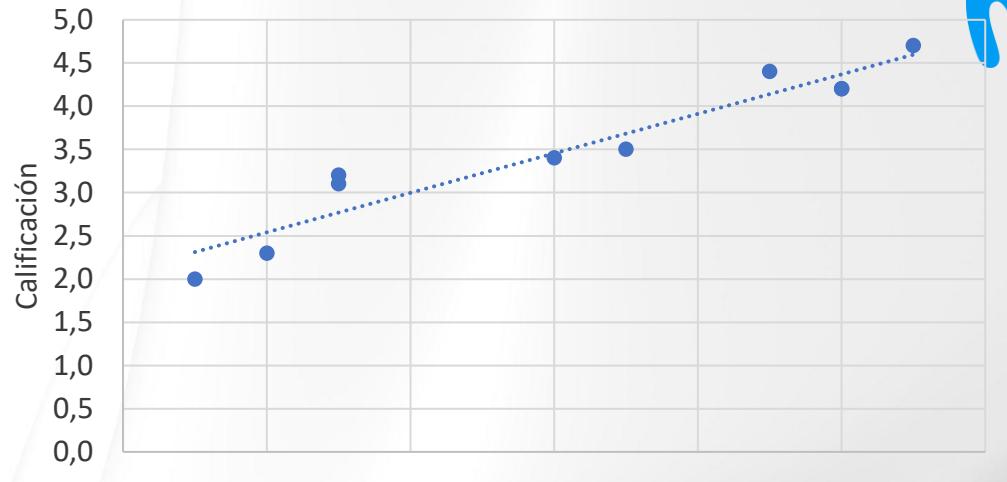
	Masculino	Femenino	Total
0 - 18 años	10	16	26
18 - 40 años	49	22	71
40 - 65 años	22	32	54
Más de 65 años	5	17	22
Total	86	87	173

	Masculino	Femenino	Total
0 - 18 años	5.8%	9.2%	15.0%
18 - 40 años	28.3%	12.7%	41.0%
40 - 65 años	12.7%	18.5%	31.2%
Más de 65 años	2.9%	9.8%	12.7%
Total	49.7%	50.3%	100%

Análisis bivariado

Variables cuantitativas

- El análisis bivariado compara dos cosas para ver su relación.
 - Por ejemplo, podrías ver si a más horas de estudio corresponde una mejor nota en un examen.
- Si cuando aumenta una variable, la otra también aumenta (como más horas de estudio y mejores notas), eso se llama correlación positiva. Si pasa lo contrario, sería correlación negativa





Tratamiento de variables

Label encoding

Permite representar numéricamente una variable categórica ordinal.

¿Por qué es útil?

Los modelos de Machine Learning trabajan con variables numéricas, por tanto, se busca convertir las categorías en valores numéricos.

Categoría
Malo
Regular
Bueno
Excelente

Label Encoding

Valor
0
1
2
3

Nivel educativo
Primaria
Secundaria
Pregrado
Posgrado

Label Encoding

Valor
0
1
2
3

One hot encoding

Buscamos crear una nueva variable para cada categoría.

- Lo ideal es usarlo en variables categóricas **nominales**.

id	color
1	red
2	blue
3	green
4	blue

One hot
encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

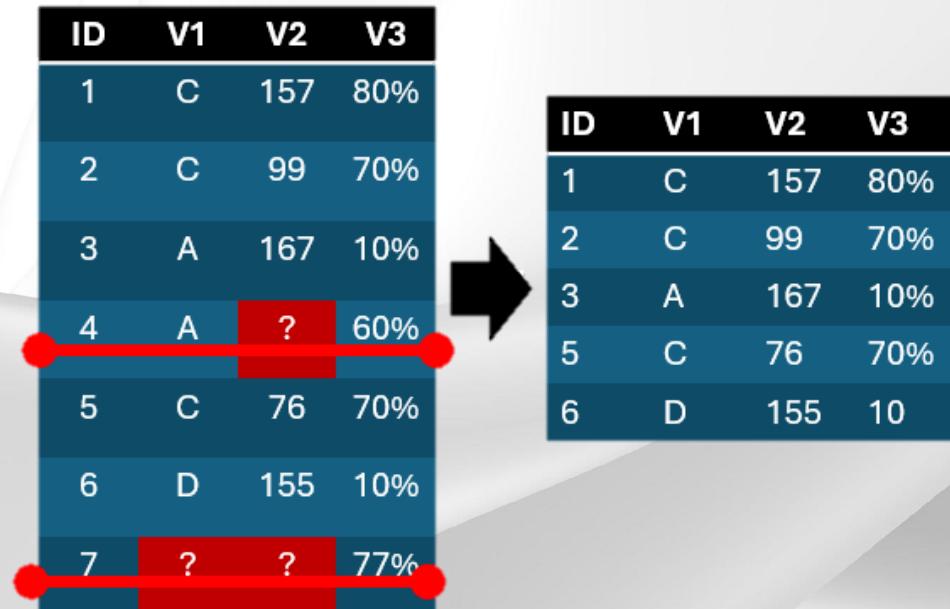
Datos Faltantes

Para plantear un modelo, lo ideal es tener completitud de la información.

Las bases de datos pueden tener datos faltantes, los cuales pueden representar problemas para analizar nuestra información.

Algunos modelos de Machine Learning son incapaces de trabajar con información faltante.

¡Podemos solucionarlo con diferentes técnicas!



The diagram illustrates a data transformation process. On the left, a table shows a dataset with missing values (indicated by question marks). A red horizontal line connects the fourth row of the first table to the seventh row of the second table. A large black arrow points from the first table to the second table, indicating the flow of data from the original state to the cleaned state.

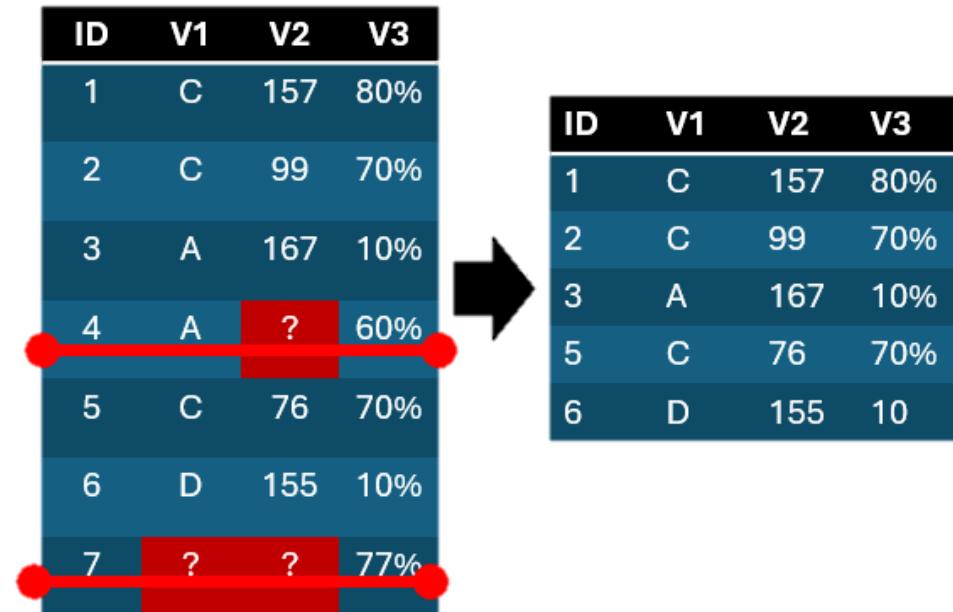
ID	V1	V2	V3
1	C	157	80%
2	C	99	70%
3	A	167	10%
4	A	?	60%
5	C	76	70%
6	D	155	10%
7	?	?	77%

ID	V1	V2	V3
1	C	157	80%
2	C	99	70%
3	A	167	10%
4	C	76	70%
5	D	155	10%

Datos Faltantes

1. Descarte de Datos

- **Eliminar** los registros que contengan datos faltantes.



The diagram illustrates the process of data cleaning. On the left, a table contains 7 rows of data. Rows 4 and 7 are highlighted with red boxes around their V2 and V3 columns respectively, indicating they contain missing values. An arrow points from this table to the right, where a second table shows the result after cleaning: rows 4 and 7 have been removed, leaving only 5 valid rows.

ID	V1	V2	V3
1	C	157	80%
2	C	99	70%
3	A	167	10%
4	A	?	60%
5	C	76	70%
6	D	155	10%
7	?	?	77%

ID	V1	V2	V3
1	C	157	80%
2	C	99	70%
3	A	167	10%
5	C	76	70%
6	D	155	10%

Datos Faltantes

2. Imputación de Datos

- Podemos llenar los datos vacíos con el valor promedio de los datos.
- Se pueden usar otras medidas para completar los datos.

Datos por Imputar

ID	V1	V2	V3
1	25	?	50
2	27	3	?
3	29	5	110
4	31	7	140
5	33	9	170
6	?	11	200

Media = 29

Resultado Imputación

ID	V1	V2	V3
1	25	7	50
2	27	3	134
3	29	5	110
4	31	7	140
5	33	9	170
6	29	11	200



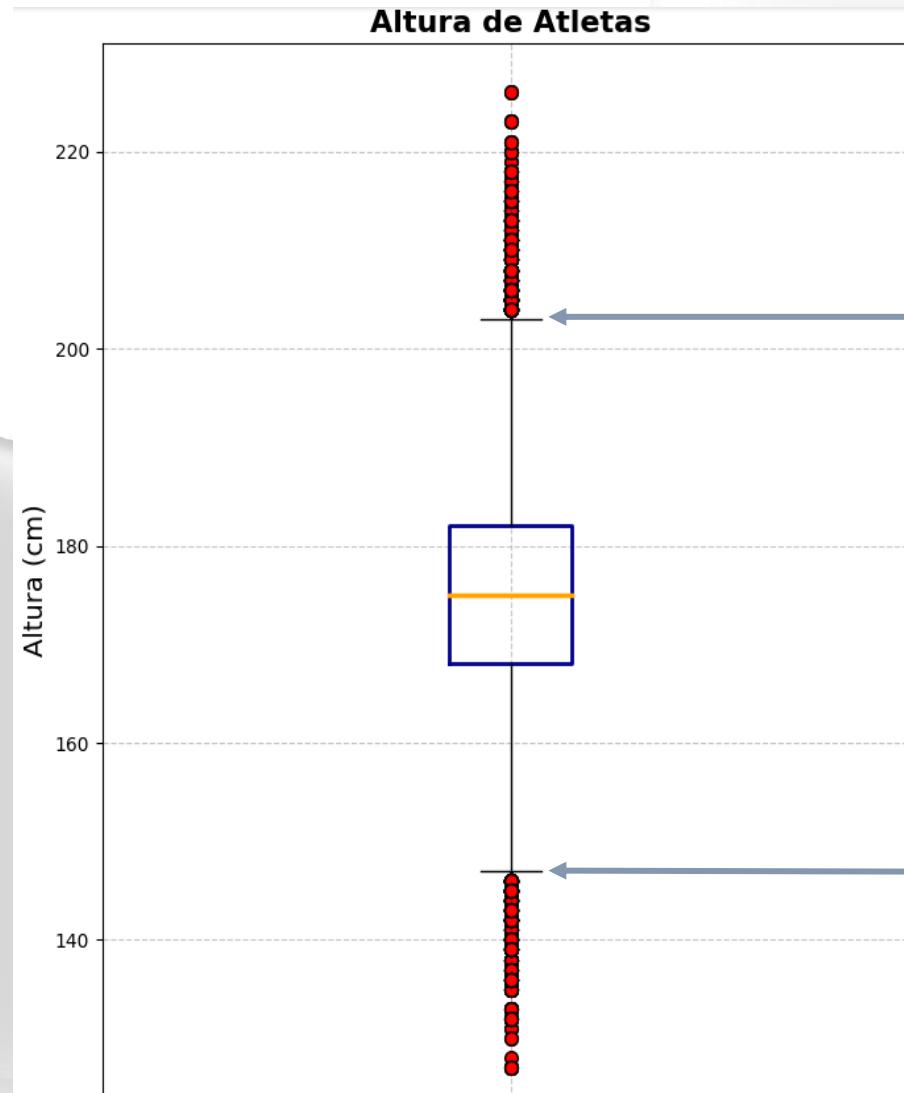
Outliers



Outliers

¿En qué consiste?

Son todos esos **valores** que se **alejan de la tendencia general** del conjunto de valores

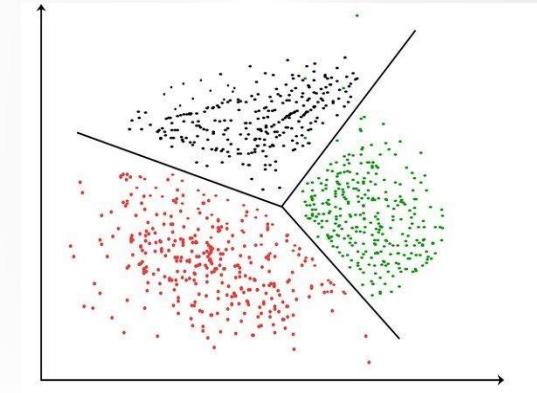


Importancia de los outliers



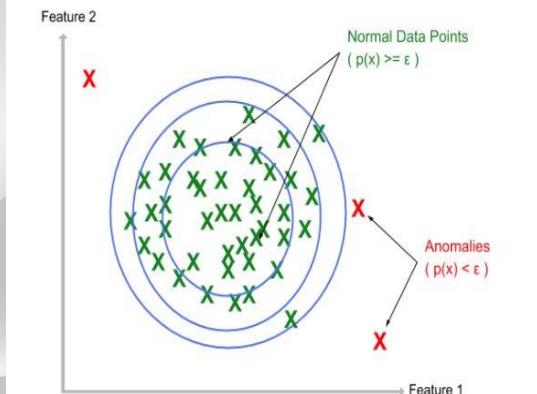
Influencia en las **métricas estadísticas**

Impacto en la **toma de decisiones**



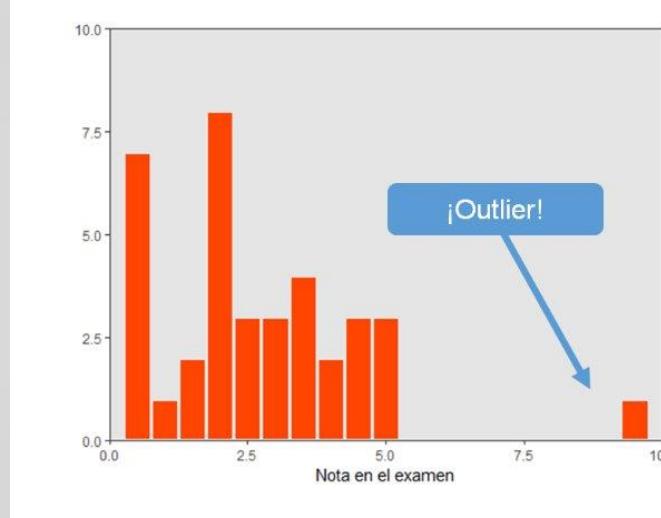
Identificación de **patrones y tendencias**

Detección de **anomalías y fraudes**

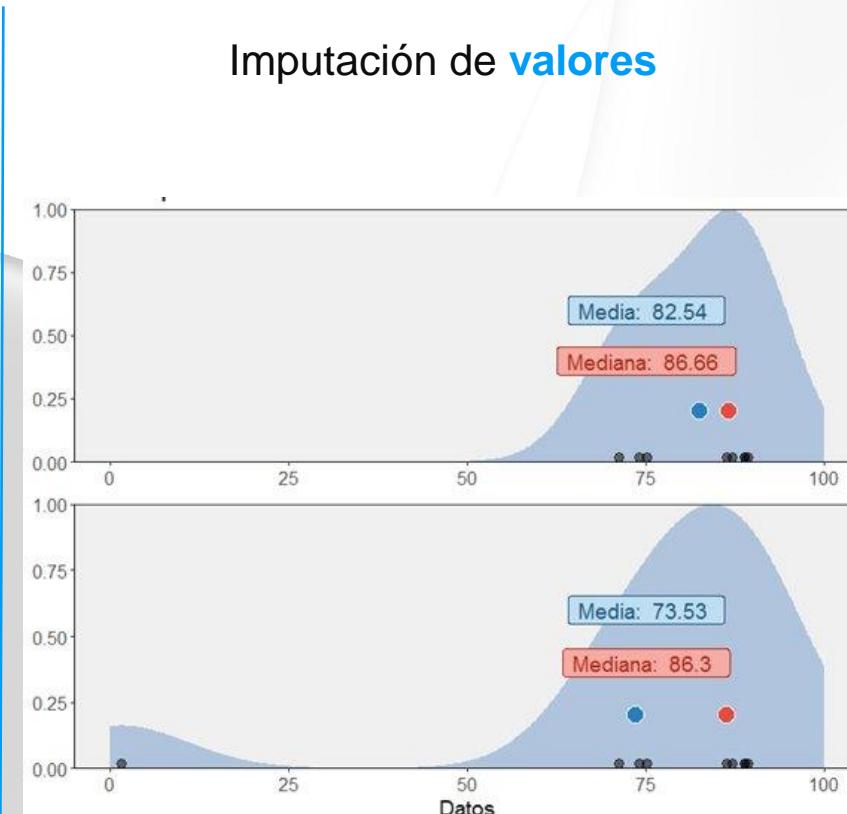


¿Qué hacer con los outliers?

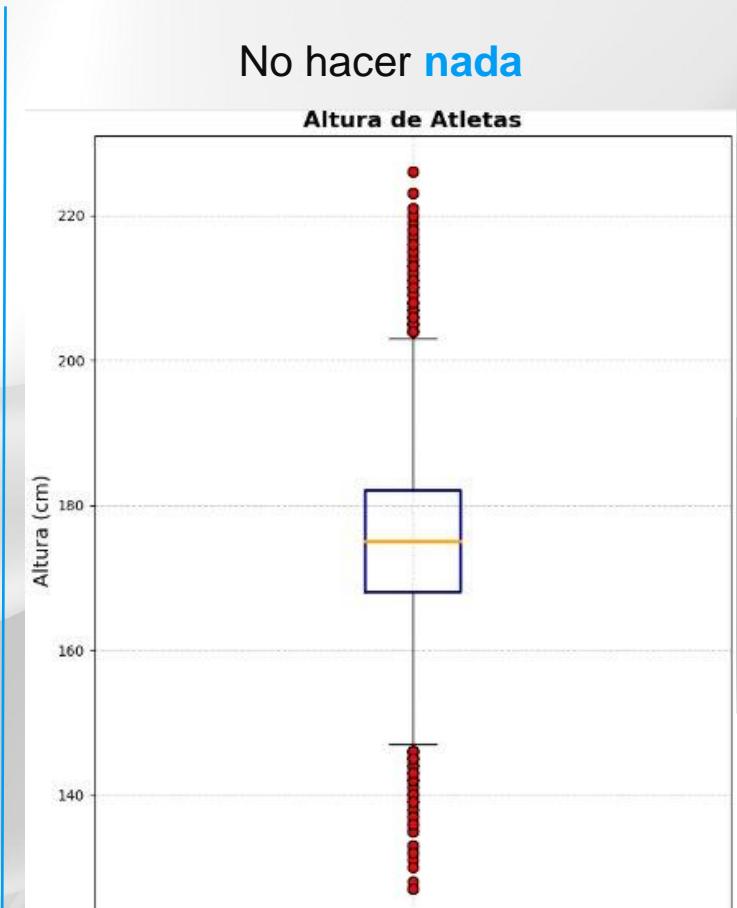
Eliminar **trimming**



Imputación de **valores**



No hacer **nada**





Agriculture



Visualización de datos



Visualización de datos

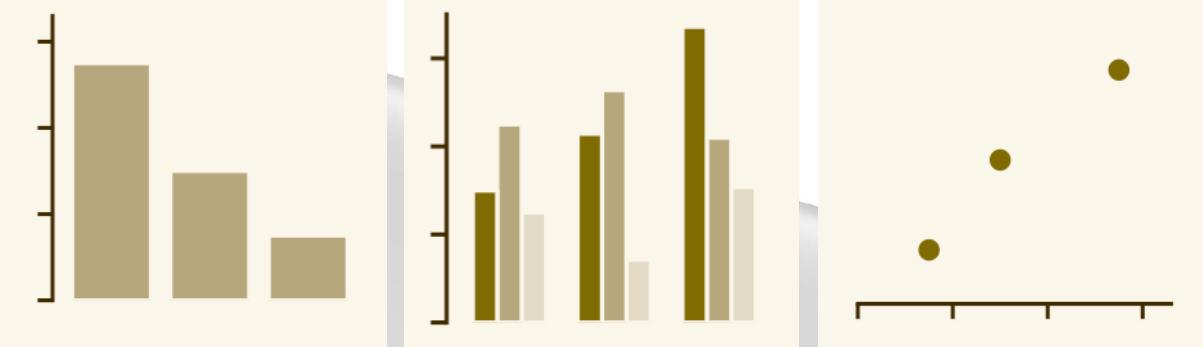
¿En qué consiste?

- Herramienta esencial que permite transformar datos complejos en representaciones **visuales fáciles de interpretar**
- A través de gráficos, diagramas y otros elementos visuales, es posible **identificar patrones, tendencias, y relaciones en los datos** que de otra manera podrían pasar desapercibidos.

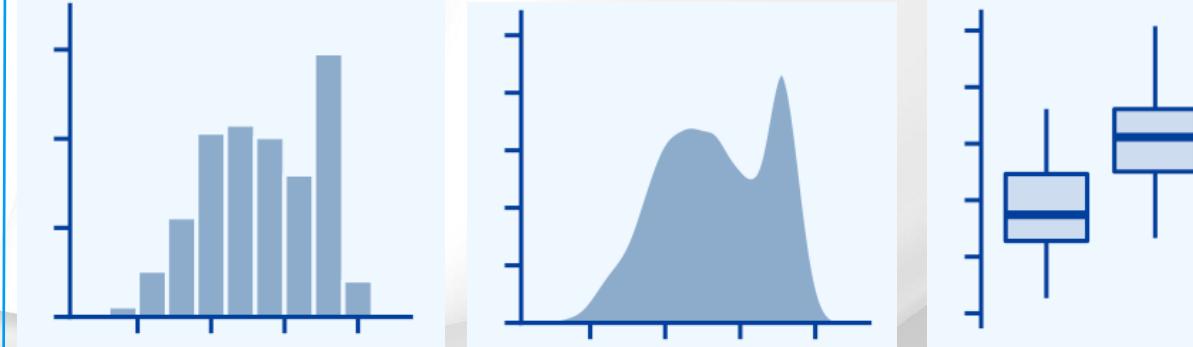


Tipos de gráficas

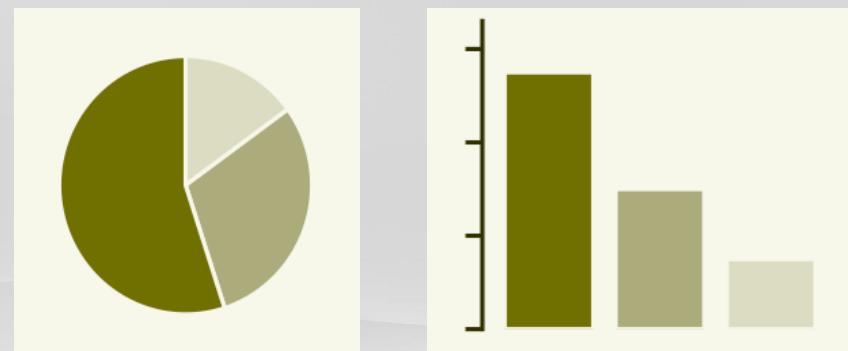
Cantidades



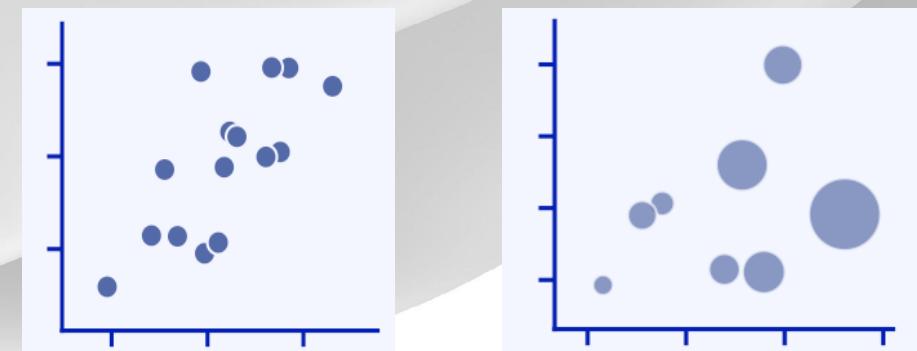
Distribuciones



Proporciones



Relacionales





Gráficas para cantidades

Gráfico de barras

Es una representación gráfica que busca mostrar la **cantidad** presente por **categorías**



Gráfico de barras

Frutas disponibles en tienda	
Fruta	Cantidad
Manzana	12
Pera	8
Guayaba	6
Fresa	10
Banano	6

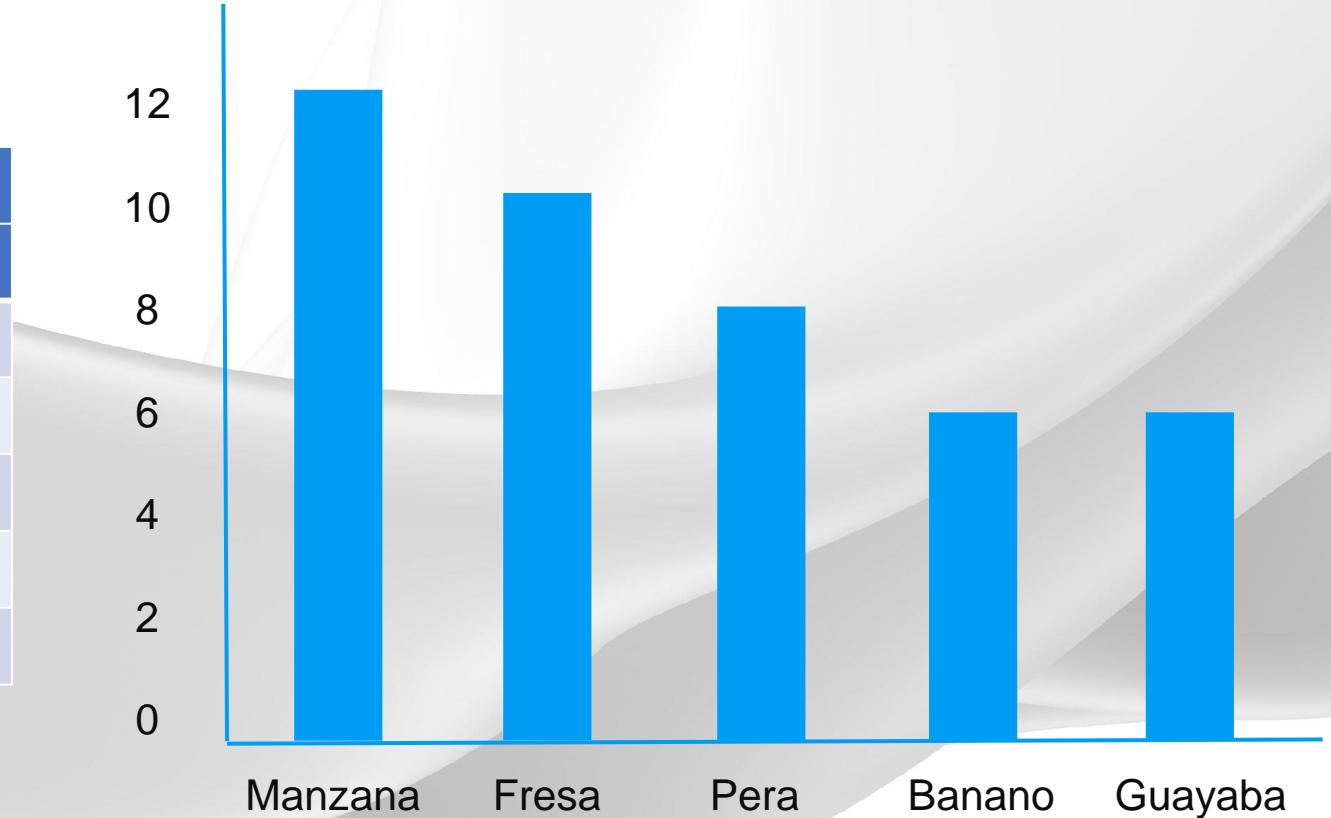


Gráfico de barras

Frutas disponibles en tienda	
Fruta	Cantidad
Manzana	12
Pera	8
Guayaba	6
Fresa	10
Banano	6

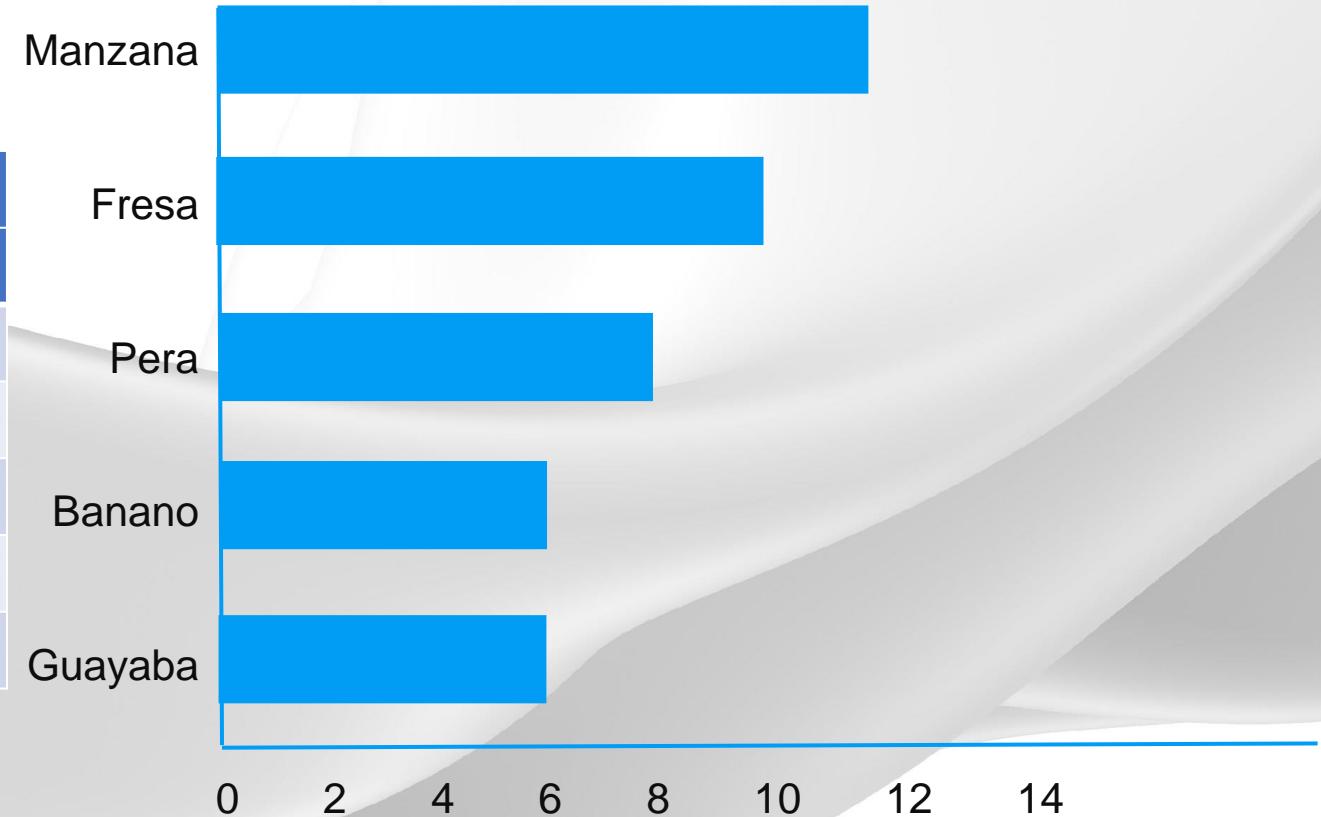


Gráfico de barras agrupado

Es una representación gráfica que busca mostrar la **cantidad** presente en **una categoría** y cómo está distribuido en **otra categoría**

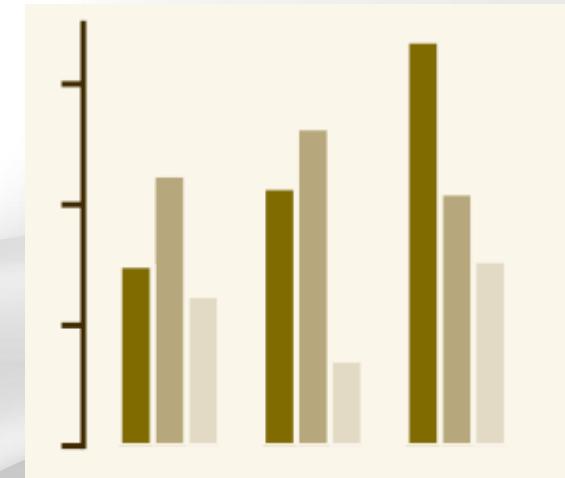
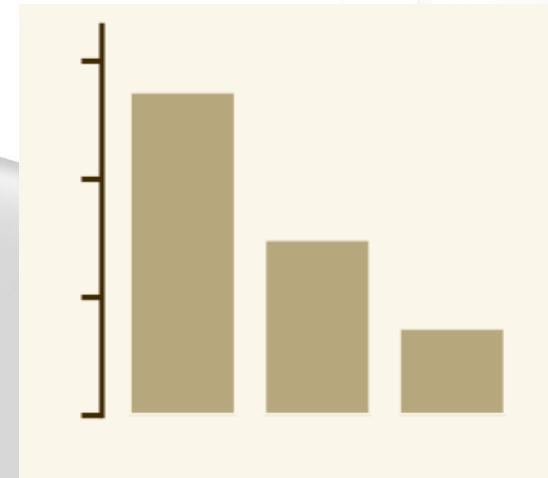
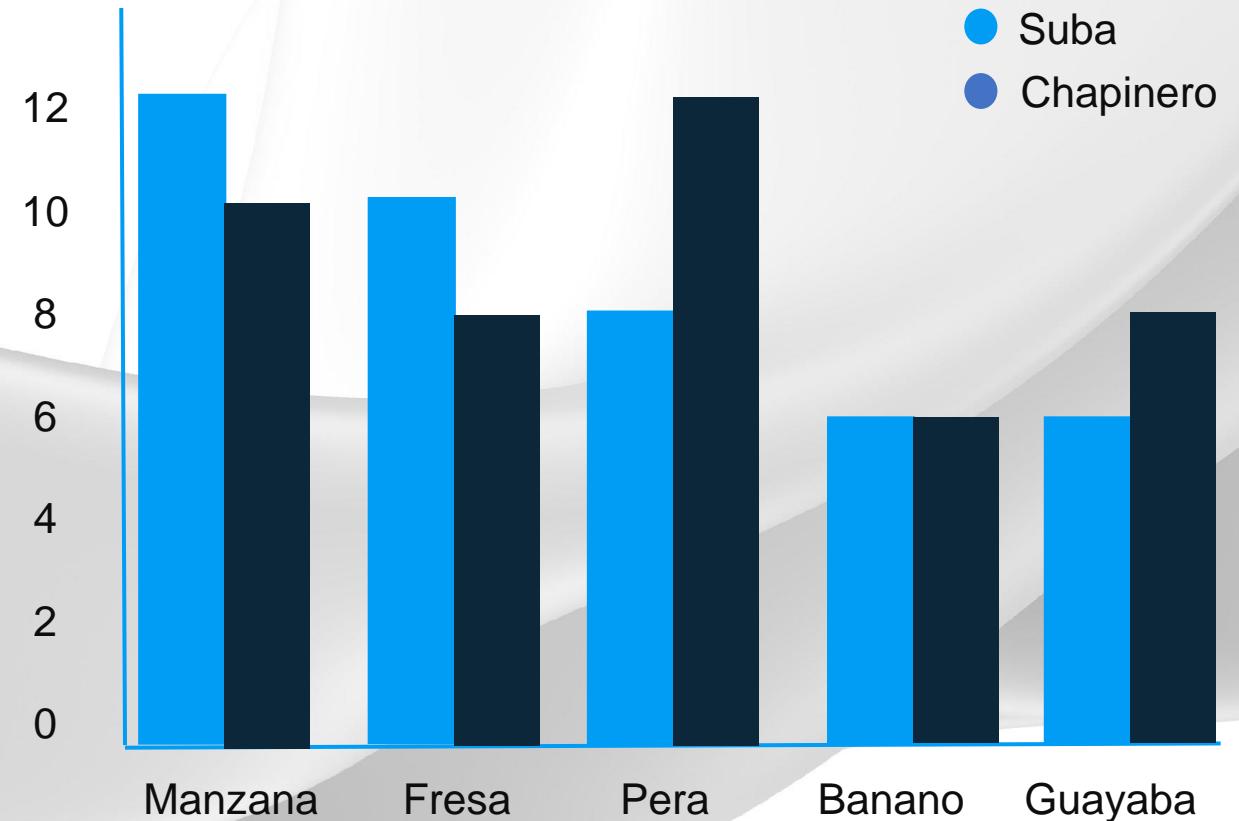


Gráfico de barras agrupado

Frutas disponibles por tienda		
Fruta	Suba	Chapinero
Manzana	12	10
Pera	8	12
Guayaba	6	8
Fresa	10	8
Banano	6	6

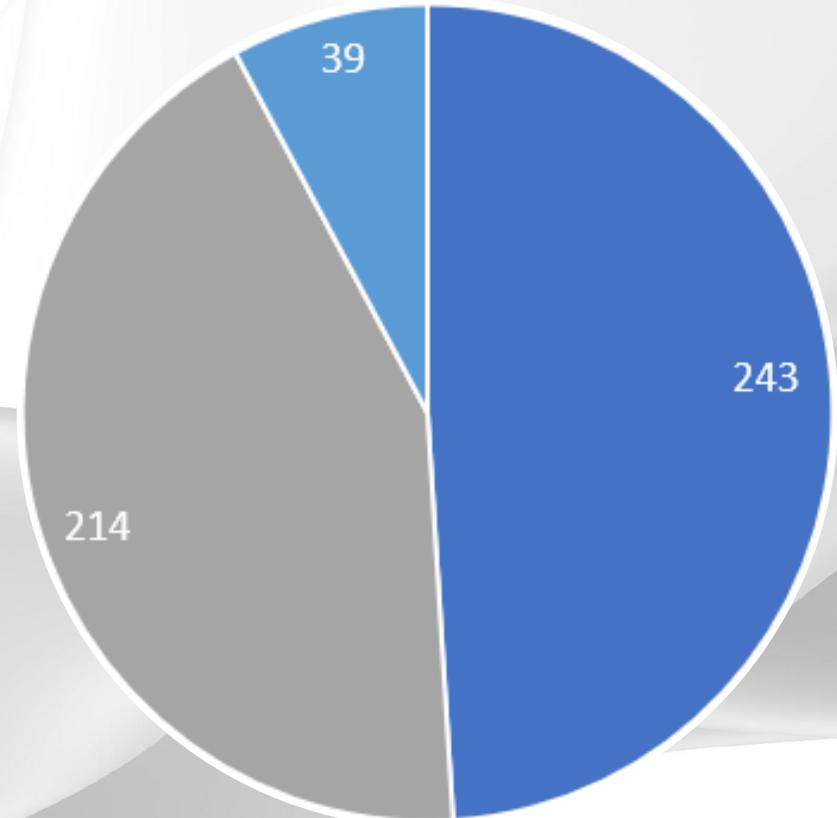




Gráficas para proporciones

Gráfico de torta

Es una representación gráfica que busca mostrar la **proporción** presente en una **categoría**

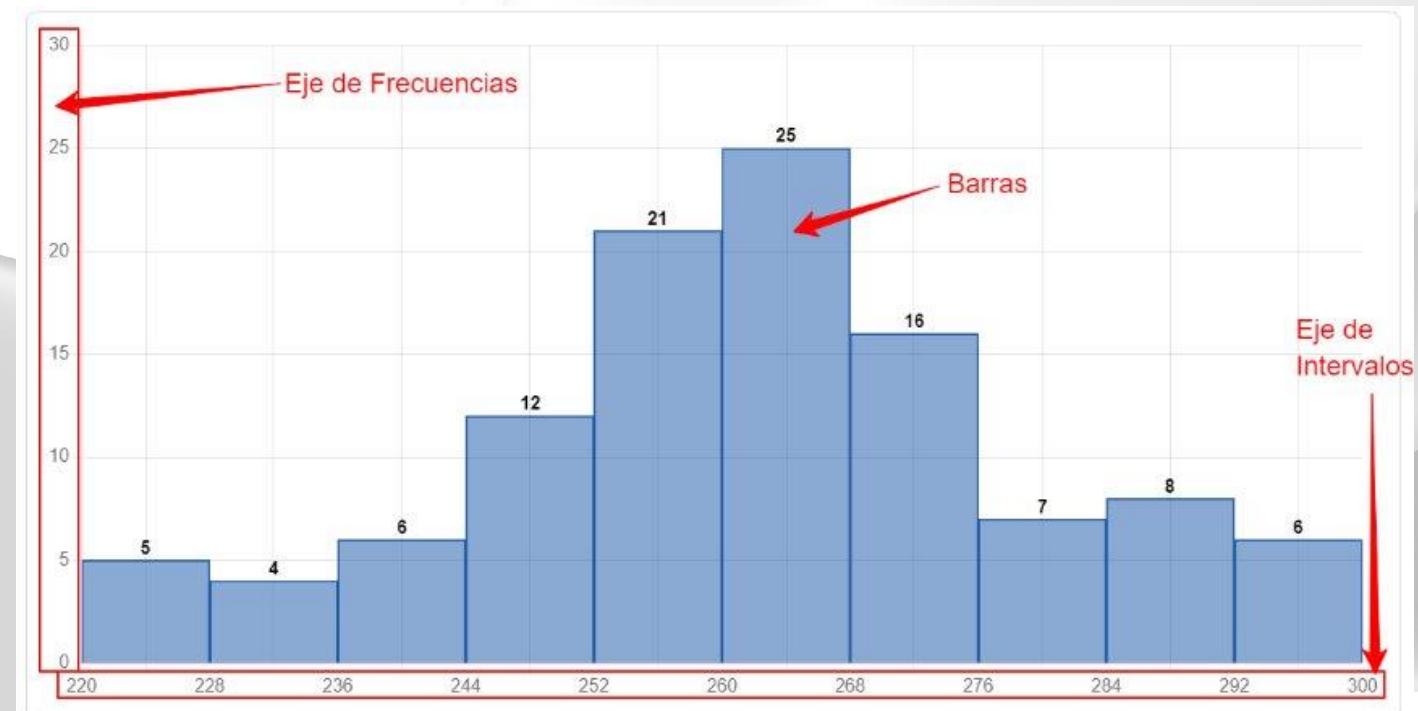




Gráficas para distribuciones

Histograma

Es una representación gráfica de un conjunto de datos que muestra la **frecuencia con la que se presenta un rango de valores**

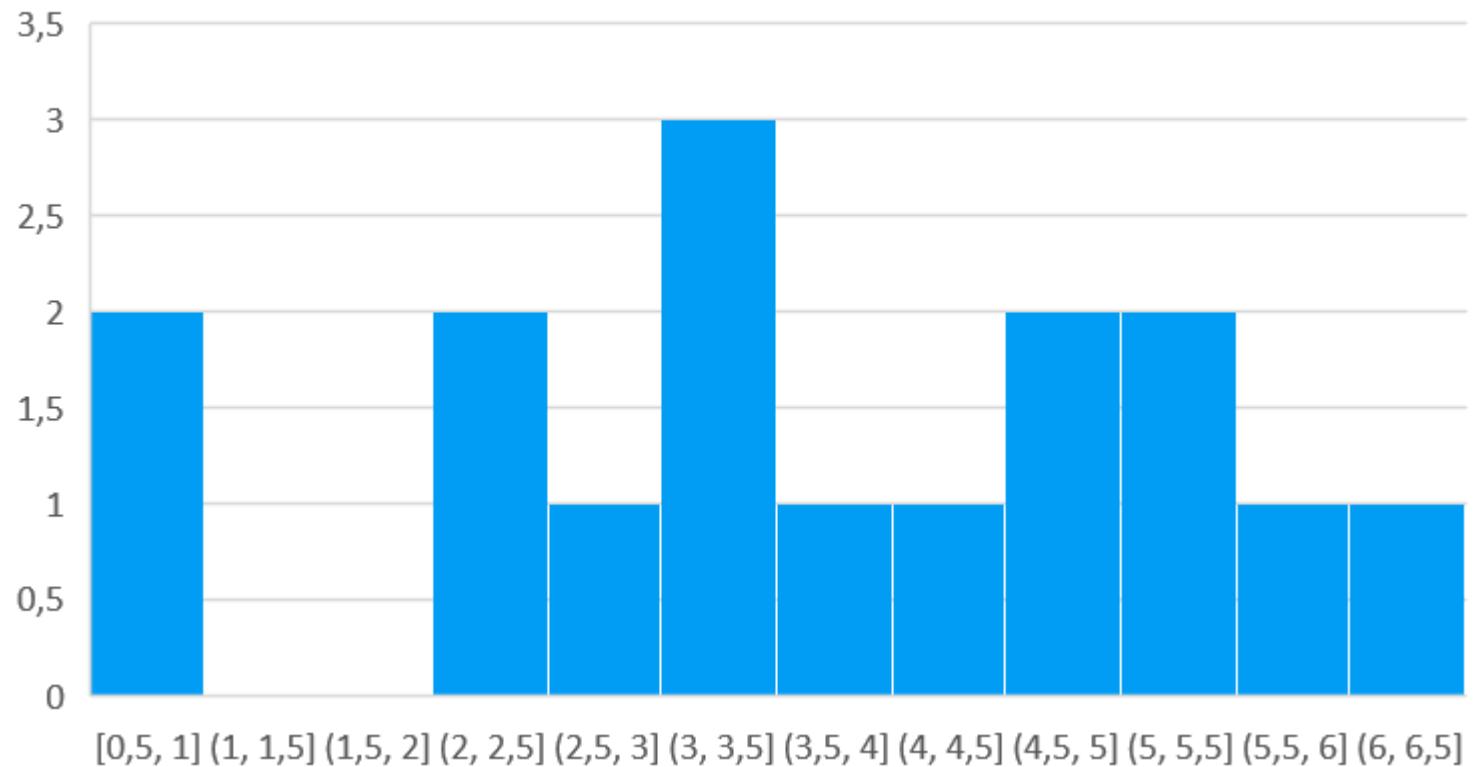


Histograma

Cantidad de vasos de agua en promedio

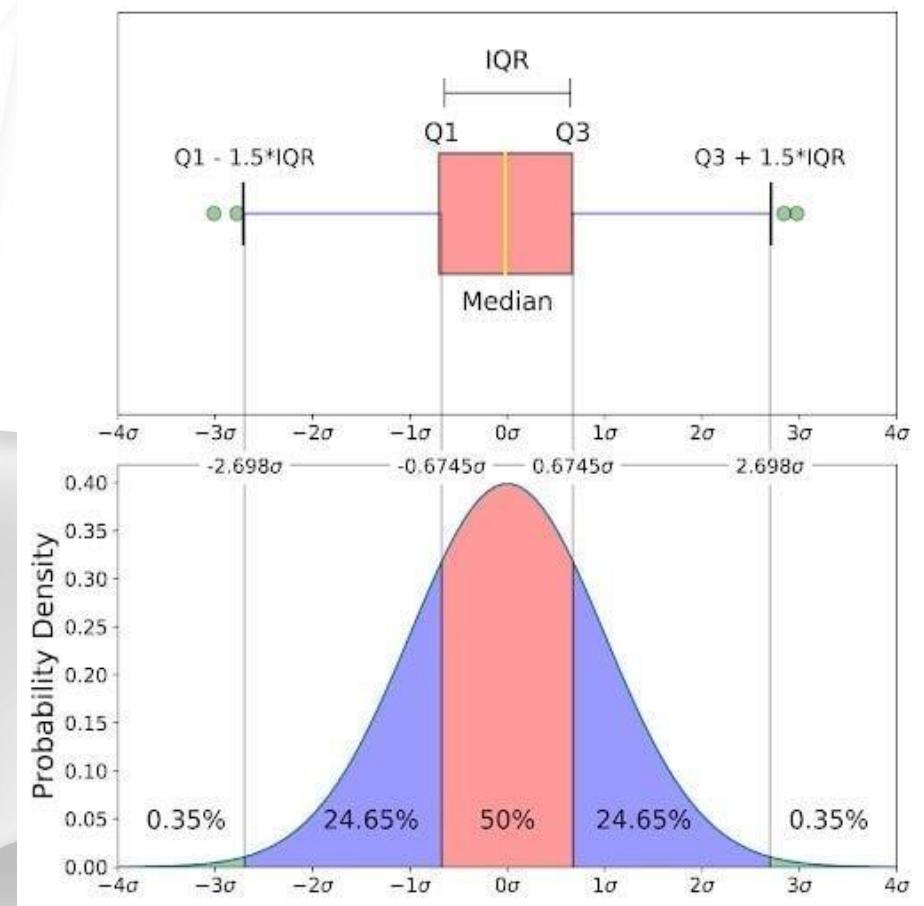
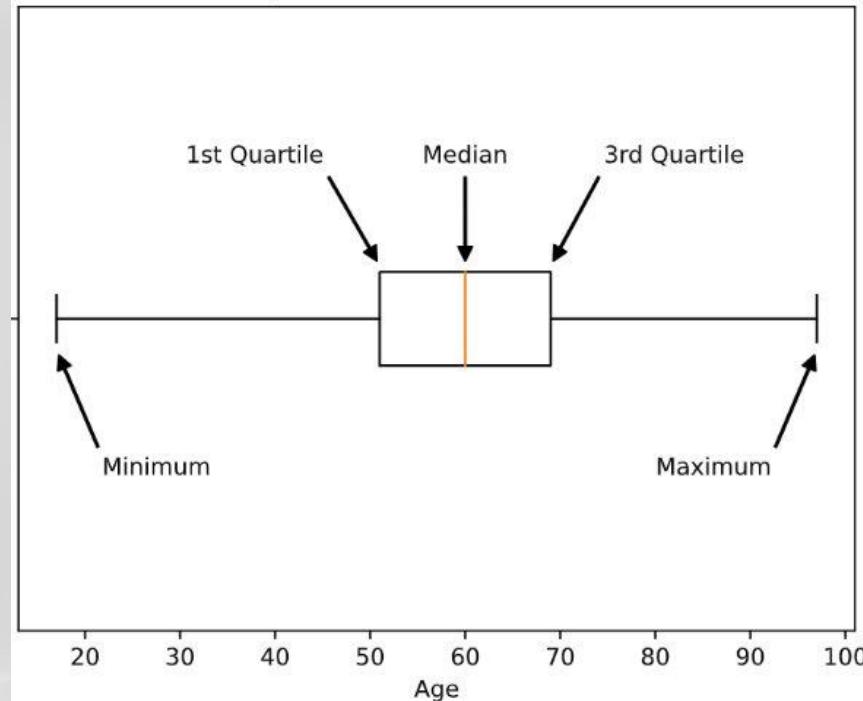
0,5
0,7
2,1
2,2
2,9
3,2
3,2
3,3
3,7
4,5
4,6
4,8
5,2
5,3
5,8
6,2

Histograma



Boxplot

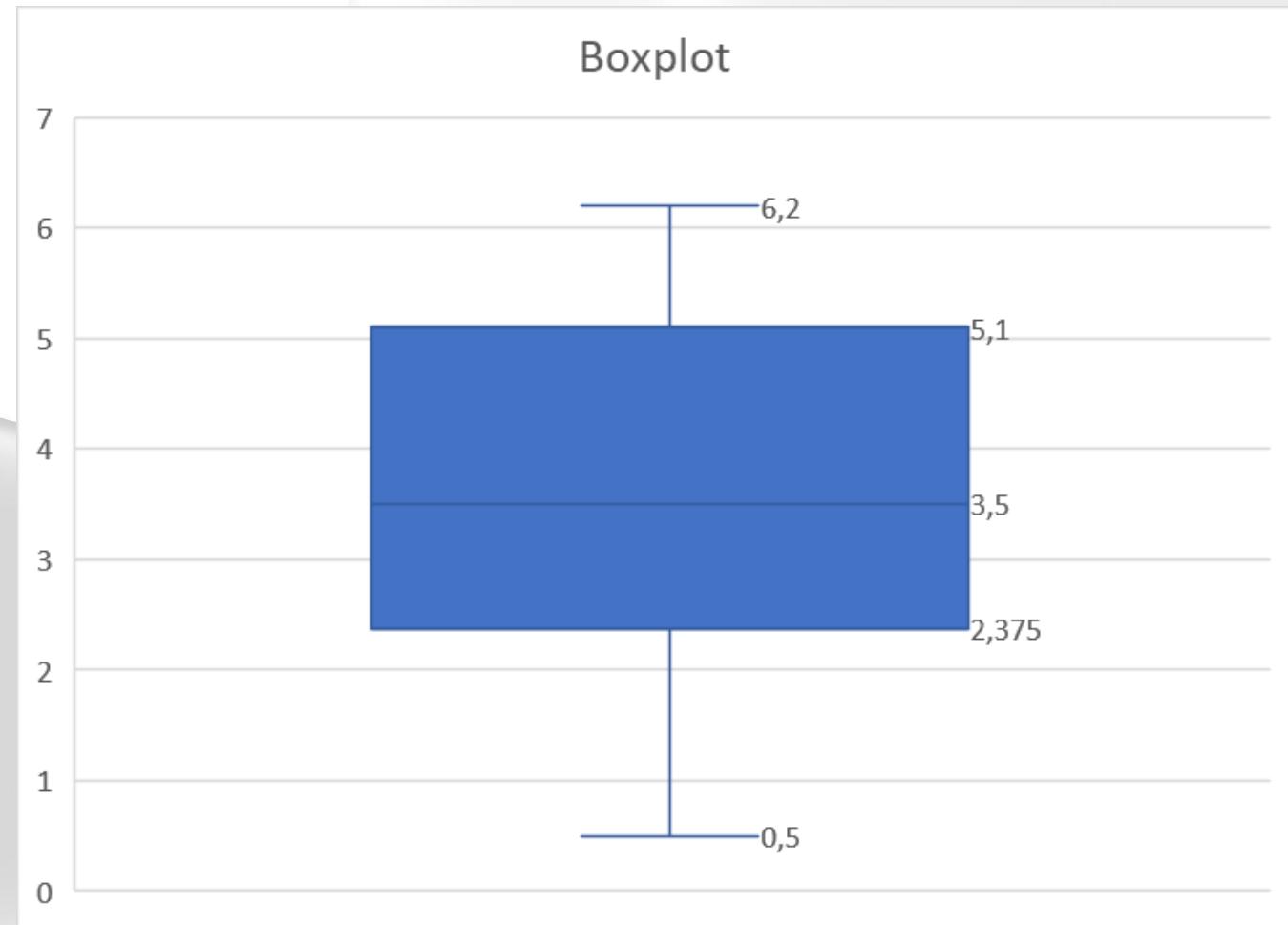
La utilizamos para identificar rápidamente el **rango** donde se encuentran la mayoría de **datos** y si hay **outliers**



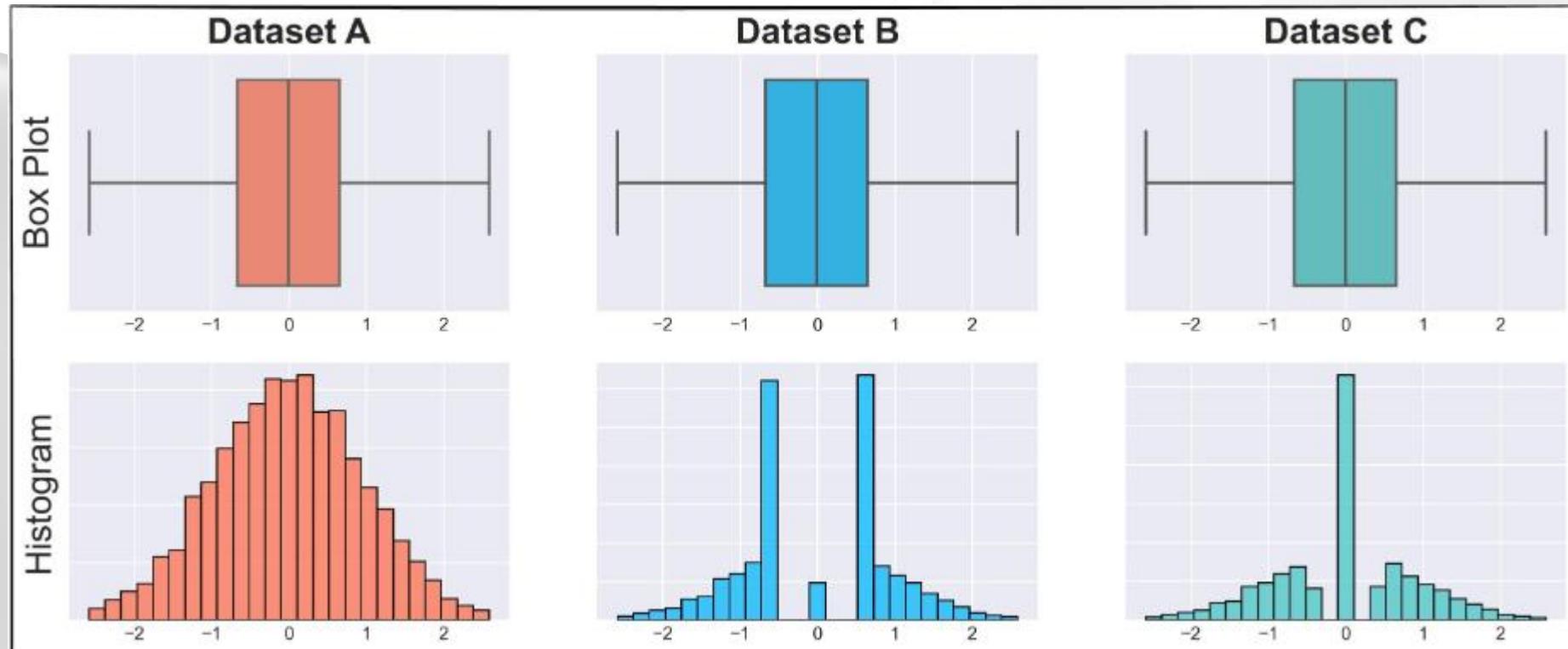
Boxplot

Cantidad de vasos de agua en promedio

0,5
0,7
2,1
2,2
2,9
3,2
3,2
3,3
3,7
4,5
4,6
4,8
5,2
5,3
5,8
6,2



Diferencia entre gráficos

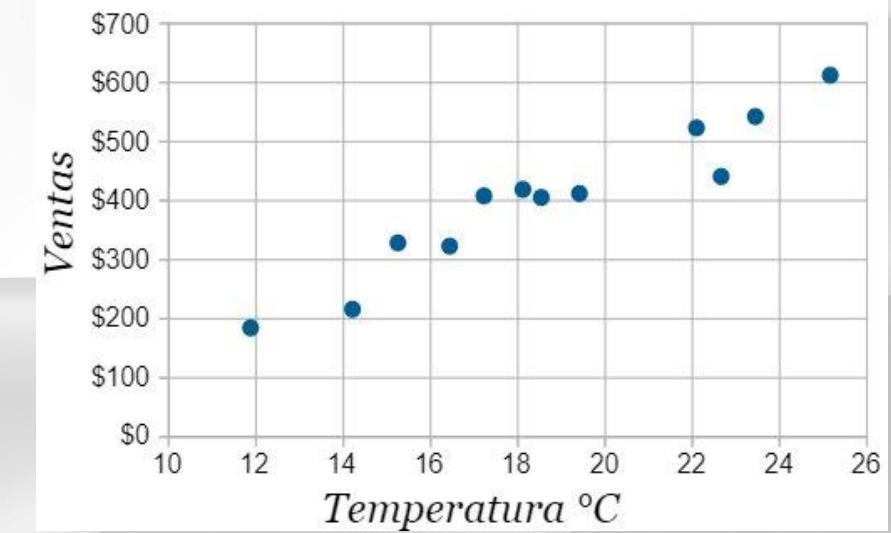
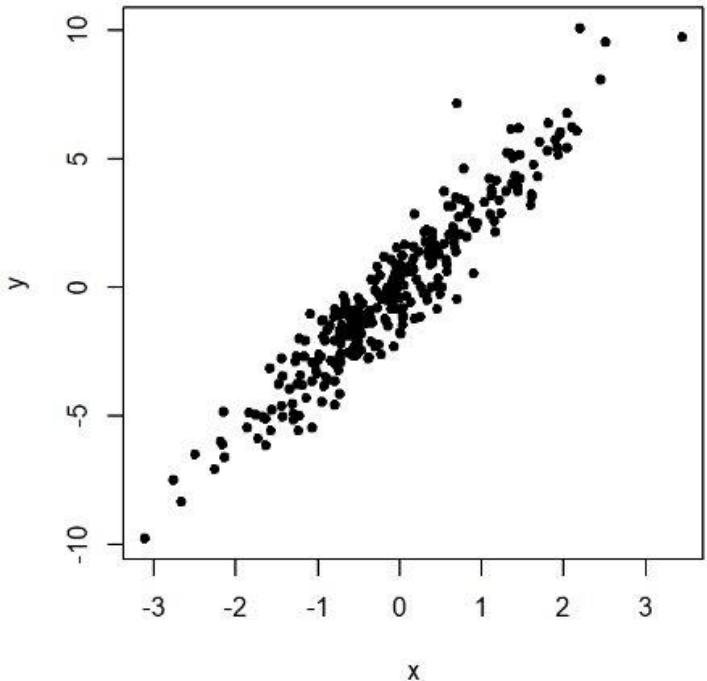




Gráficas relacionales

Gráficas de dispersión

Este gráfico lo utilizamos para **identificar la relación lineal** entre **dos variables**





Matriz de correlación

La matriz de correlación es otra manera de ver la **correlación** que hay entre **dos o más variables**

Matriz de Correlación			
medal	height	weight	
medal	1.00	0.08	0.08
height	0.08	1.00	0.80
weight	0.08	0.80	1.00

	ventas	gastos_marketing	clientes_nuevos	satisfaccion_cliente
ventas	1.00	0.61	0.72	-0.57
gastos_marketing	0.61	1.00	0.43	-0.37
clientes_nuevos	0.72	0.43	1.00	-0.39
satisfaccion_cliente	-0.57	-0.37	-0.39	1.00



¿Qué es machine learning?

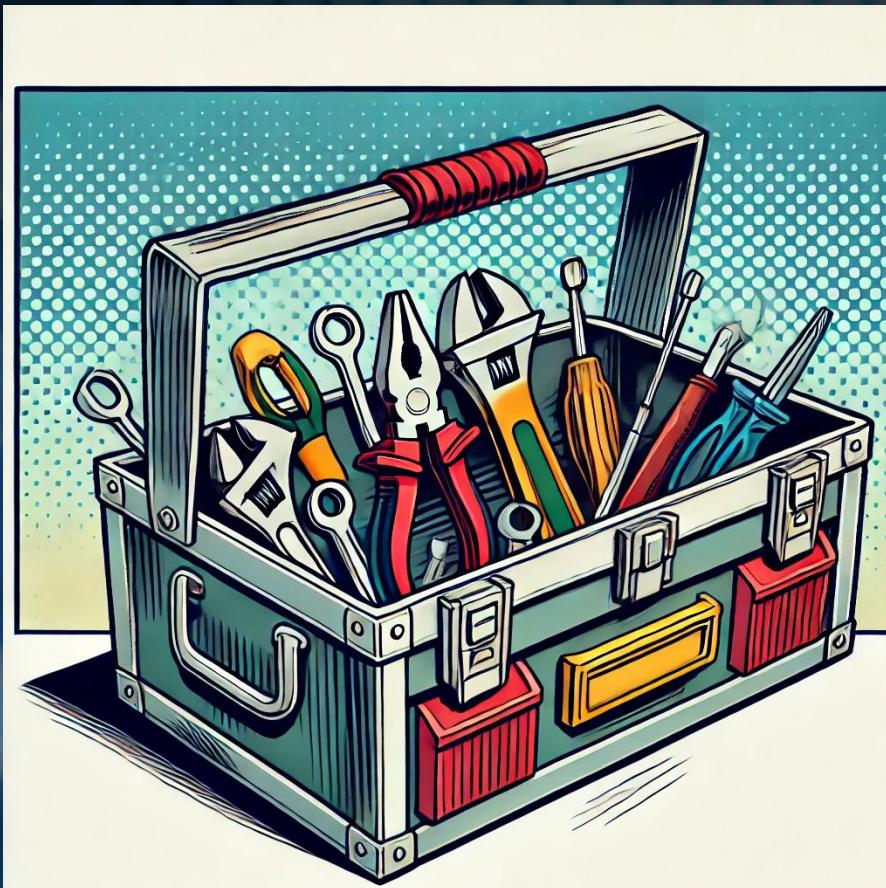
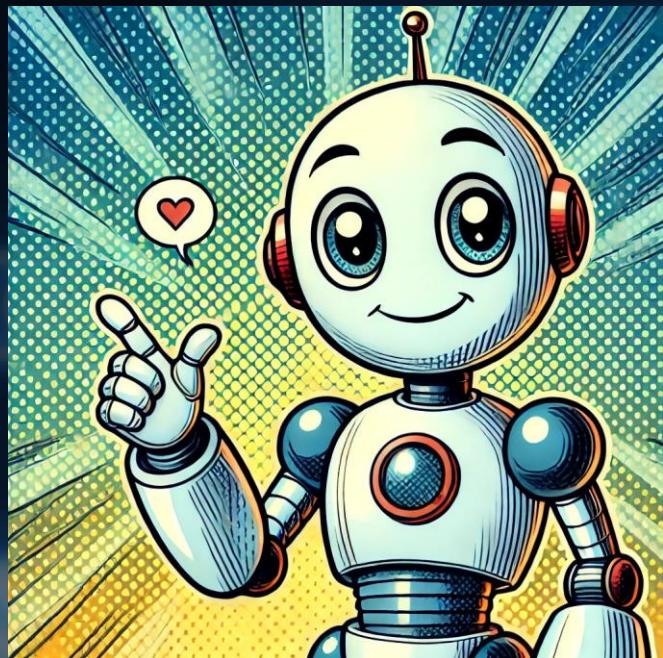




**Queremos escucharlos:
¿Qué es machine learning?,
para ustedes**



Queremos escucharlos:
¿Qué es machine learning?, para ustedes



¿Qué es machine learning?

- Rama de la IA enfocada en algoritmos (modelos) que permitan al computador **aprender de los datos existentes** para predecir tendencias, patrones, resultados y comportamientos futuros

IA

Disciplina de las **ciencias de la computación** que, a través de la combinación de **algoritmos**, desarrolla máquinas que intentan replicar o imitar la **inteligencia humana**.

ML

Subconjunto de la IA, donde los **humanos** entrena n a esas máquinas para que aprendan a través de los **datos** y hagan **predicciones**.

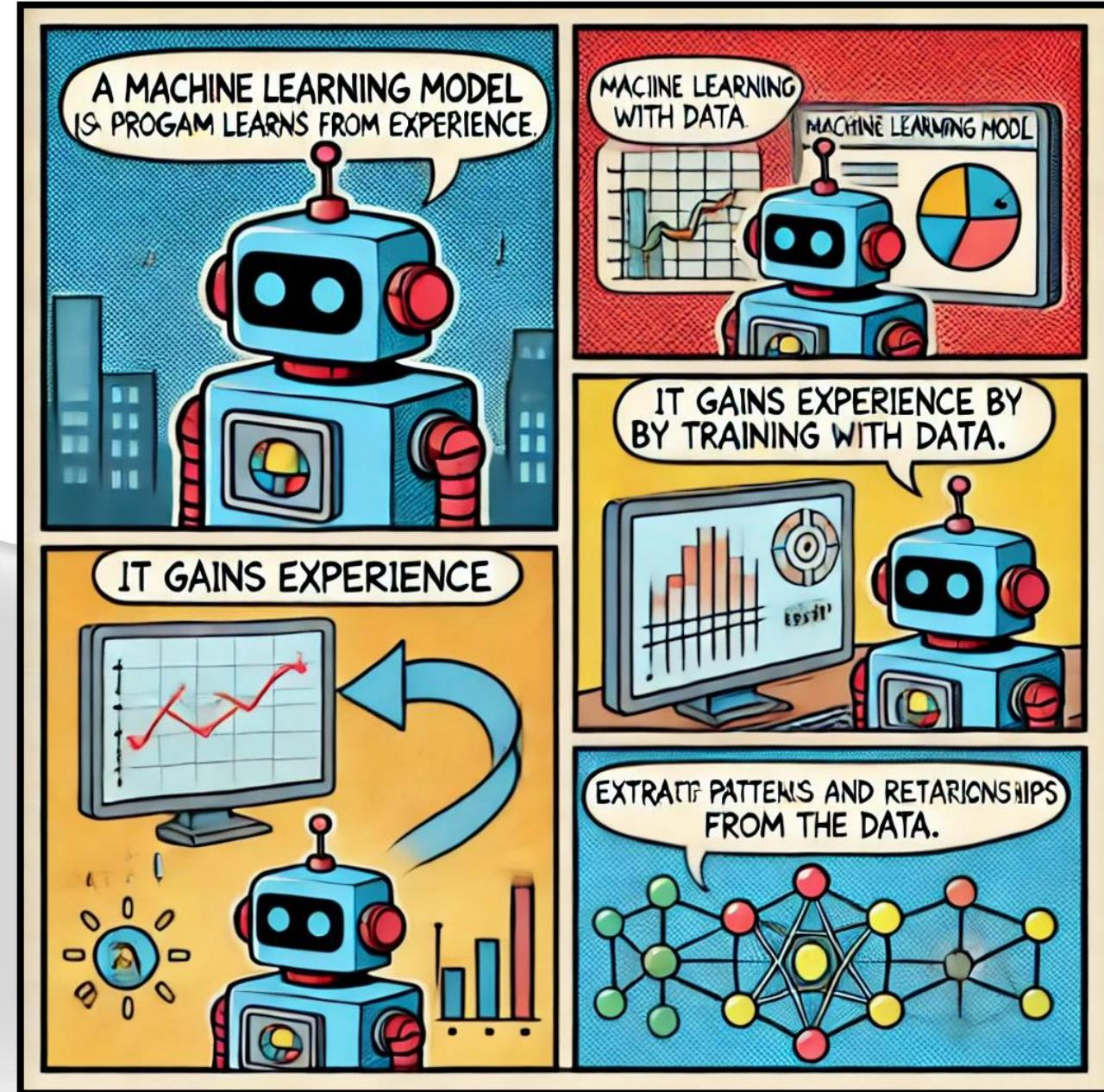
DL

Evolución del Machine Learning en la que los algoritmos están estructurados en **redes neuronales** y son capaces de **aprender por sí mismas**, **comprender esos datos** y **sacar sus propias conclusiones**.



¿Qué es modelo de machine learning?

- Un modelo de machine learning es programa el cual ayuda al computador a aprender de los datos sin explícitamente decirle que hacer
- **Aprende a través de la experiencia.**
- Obtiene la experiencia “entrenando” con datos. Extrayendo patrones y relaciones en los datos





Tipos machine learning



Tipos de Machine Learning

- Aprendizaje Supervisado: El modelo se **entrena con datos etiquetados**. Estos buscan entregar como resultado estas etiquetas

Datos Etiquetados



Cat



Dog





Tipos de Machine Learning

- Aprendizaje Supervisado: El modelo se **entrena con datos etiquetados**. Estos buscan entregar como resultado estas etiquetas
- Aprendizaje No Supervisado: El modelo se entrena con **datos sin etiquetas**. Estos buscan patrones y relaciones en los datos.





Tipos de Machine Learning

- Aprendizaje Supervisado: El modelo se **entrena con datos etiquetados**. Estos buscan entregar como resultado estas etiquetas
- Aprendizaje No Supervisado: El modelo se entrena con **datos sin etiquetas**. Estos buscan patrones y relaciones en los datos.

Datos Etiquetados



Datos NO Etiquetados





Aprendizaje No Supervisado



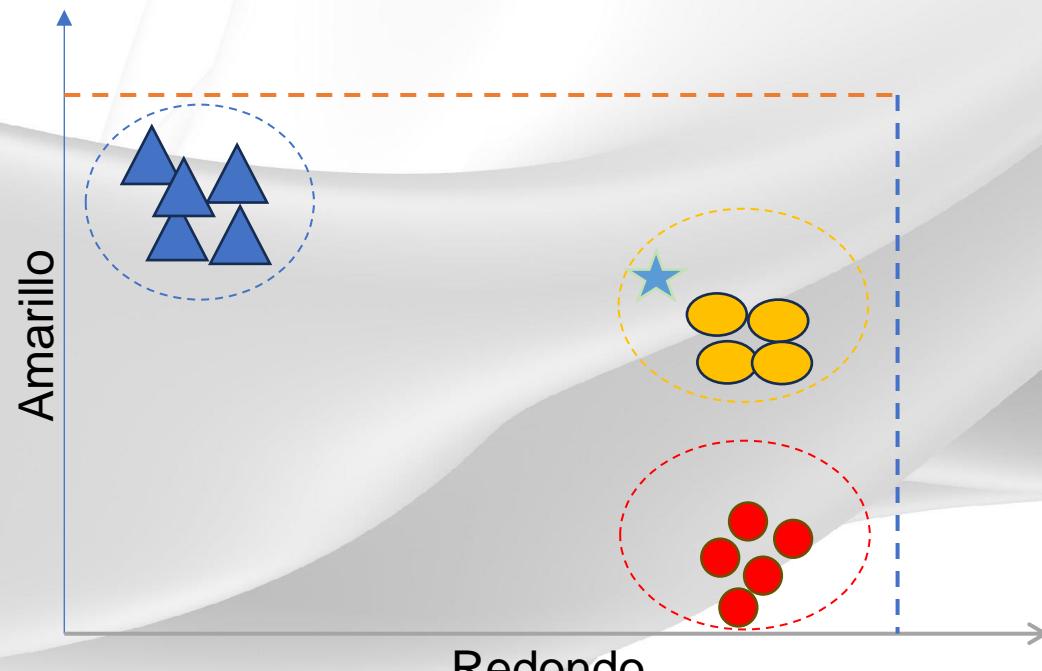
Aprendizaje NO Supervisado

- Infiere patrones a partir de un conjunto de datos sin etiquetar.
- No trata de predecir nada, solo trata de comprender patrones y agrupaciones en los datos.

Ejemplo No supervisado- Clustering

Robot clasifica diferentes frutas a partir de imágenes

No Supervisado → observa las frutas, pero NO se le indica qué fruta está viendo



Resultado: Regiones SIN Etiqueta





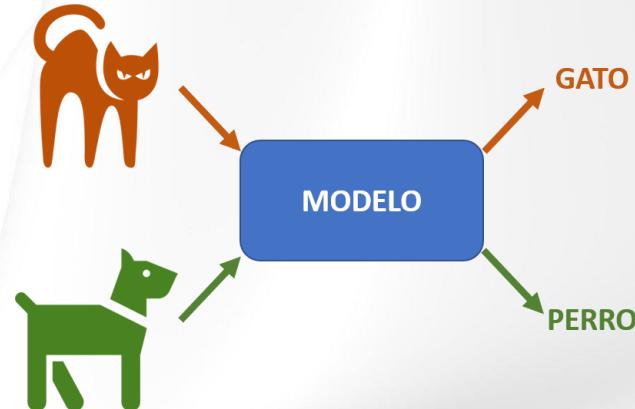
Aprendizaje Supervisado



Aprendizaje Supervisado

- Utiliza un conjunto de variables de entrada para **predecir el valor de una variable de salida** (target).
- **Aprende** acerca de un conjunto de **datos ya clasificados** para poder hacer predicciones futuras
- Hay de dos tipos de modelos **Clasificación** y **Regresión**

- Clasificación: ¿Esto es un perro o un gato?



- Regresión: ¿Qué temperatura hará mañana?





¿Cómo aprenden los modelos supervisados?



Base de datos un modelo supervisado



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Cada fila es un ejemplo diferente en los datos

Cada fila representa una persona en el Titanic



Base de datos un modelo supervisado



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

La regla general: es que entre más **datos de calidad** mejor



Base de datos un modelo supervisado



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S

Cada **columna** es una **características** diferente

Cada columna representa una característica de los pasajeros del Titanic



Base de datos un modelo supervisado



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Esta **columna**, es nuestro **target**, la variable objetivo, lo que queremos predecir

En este caso si sobrevivió el pasajero al hundimiento del Titanic

Importante: ¡Es necesario definir muy bien nuestro target!



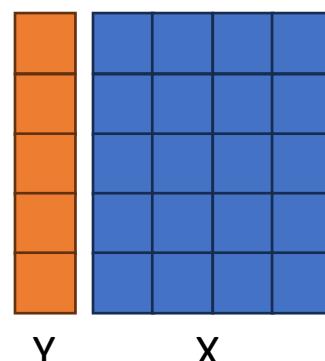
Base de datos un modelo supervisado



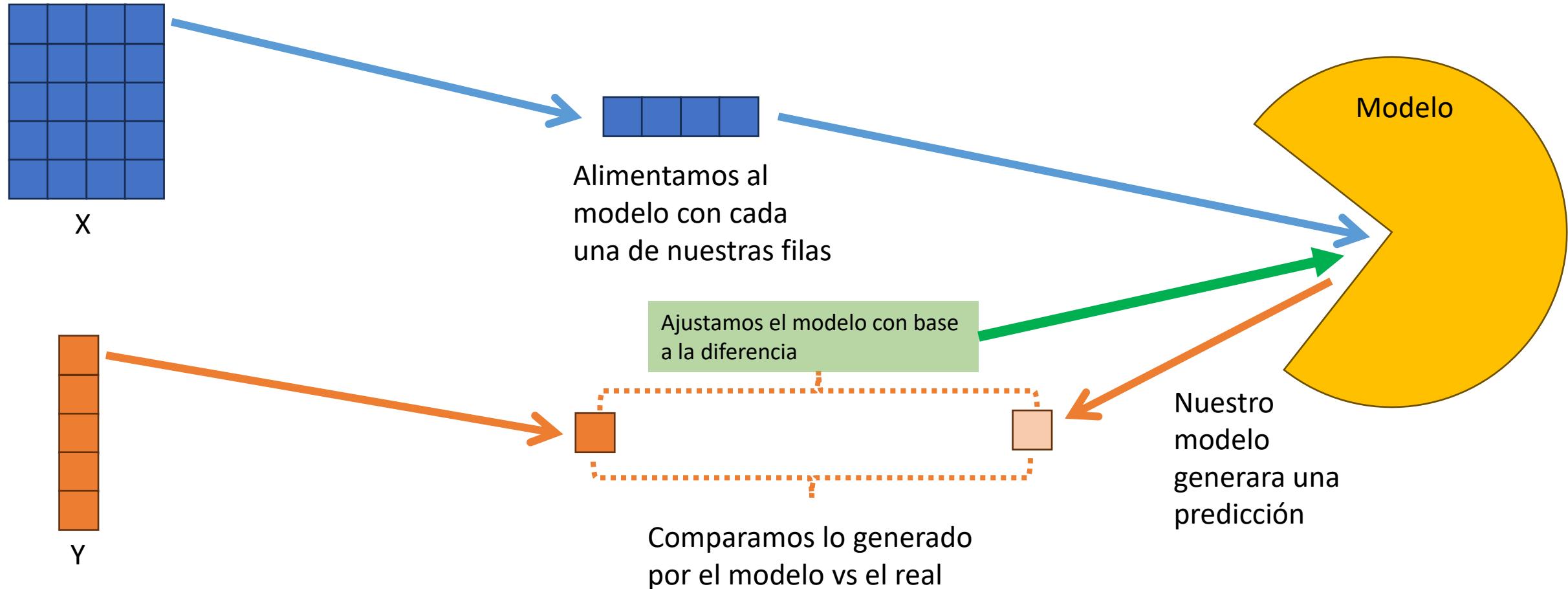
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nan	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2, 3101282	7.9250	Nan	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nan	S

Target = Y

Matriz de características = X

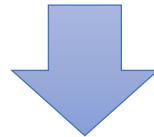


Entrenamiento de un modelo

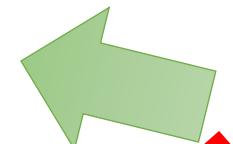


Entrenamiento de un modelo

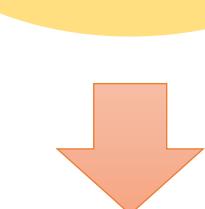
Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S



Modelo de
machine learning



Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S

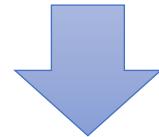


Sobrevive (1)

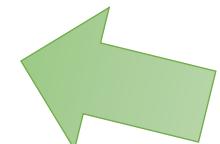
- Comparamos lo que el modelo predijo vs el valor real.
- La idea es que con cada ejemplo el modelo aprenda y se acerque más al real

Entrenamiento de un modelo

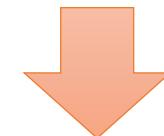
Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S



Modelo de
machine learning



Survived
0

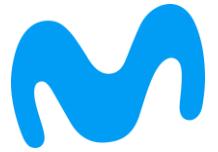


Sobrevive (0)

- Comparamos lo que el modelo predijo vs el valor real.
- La idea es que con cada ejemplo el modelo aprenda y se acerque más al real



Entrenamiento y Prueba



¿Deberíamos utilizar todos los datos que tengamos disponibles para entrenar al modelo?

¿Como se va a comportar con datos que nunca ha visto?

¿Cómo sabemos que el modelo se entrenó de manera correcta?

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Target Y

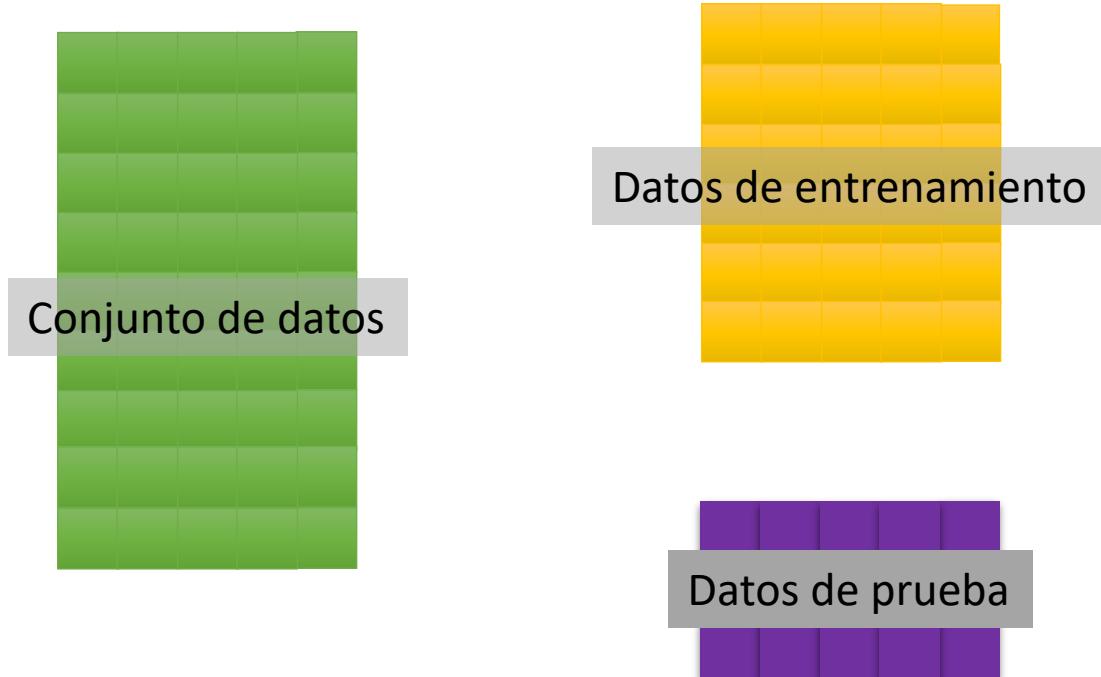
Matriz de características X

Entrenar al modelo

¿Deberíamos utilizar todos los datos que tengamos disponibles para entrenar al modelo?

¿Como se va a comportar con datos que nunca ha visto?

¿Cómo sabemos que el modelo se entrenó de manera correcta?

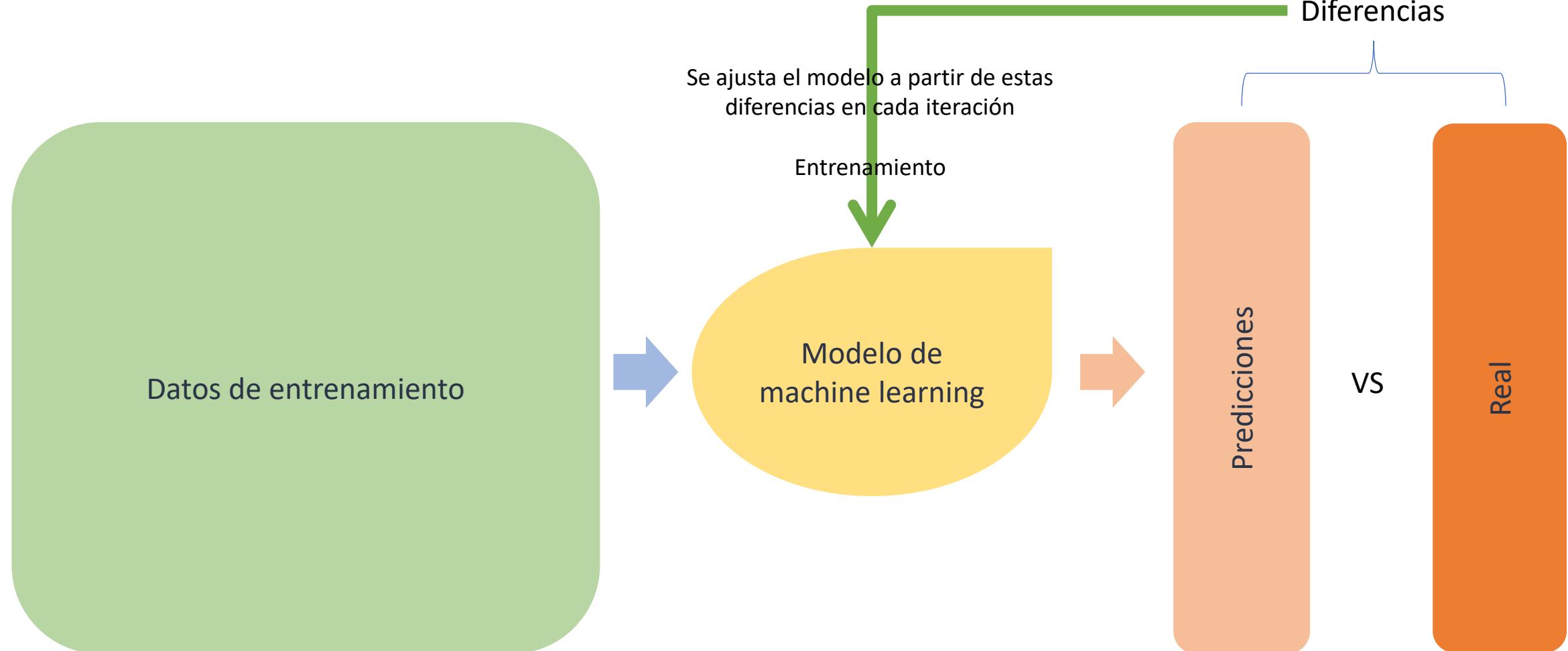


Es como cuando practicas para un examen usando algunos ejercicios.



Es como tomar el examen real después de haber practicado.

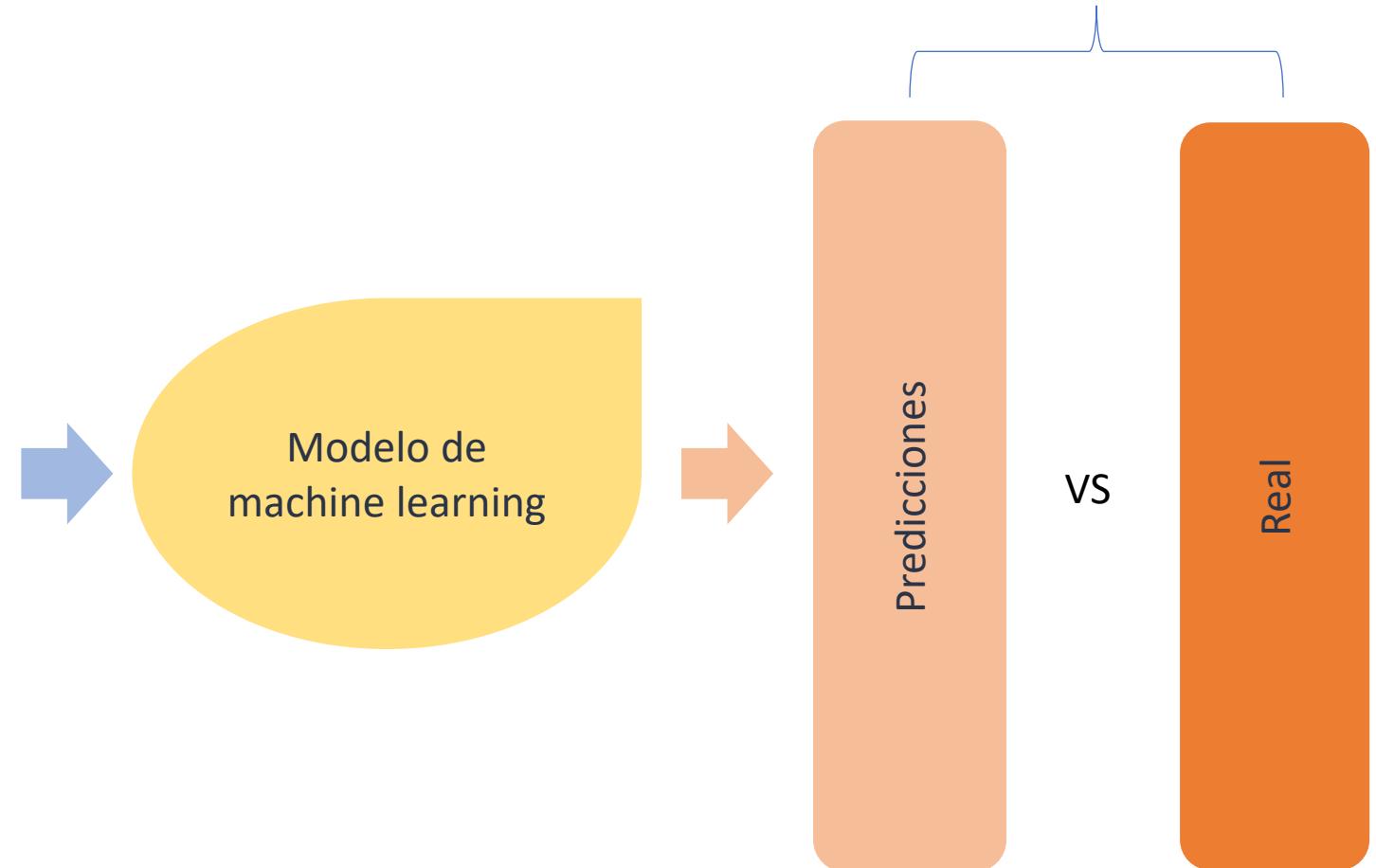
Entrenar al modelo



Entrenar al modelo



Generamos métricas de desempeño

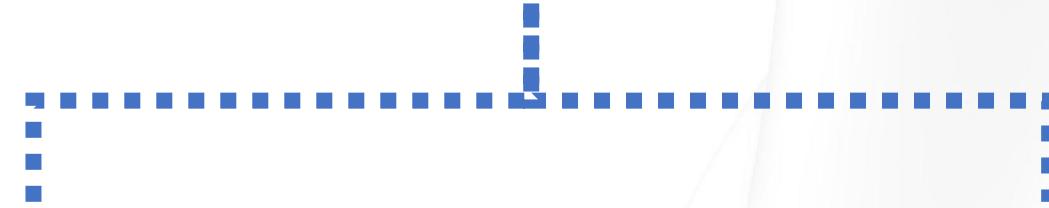




Modelos Supervisados

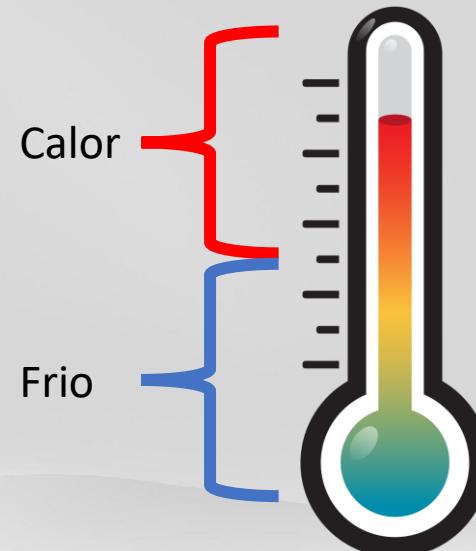


Aprendizaje Supervisado



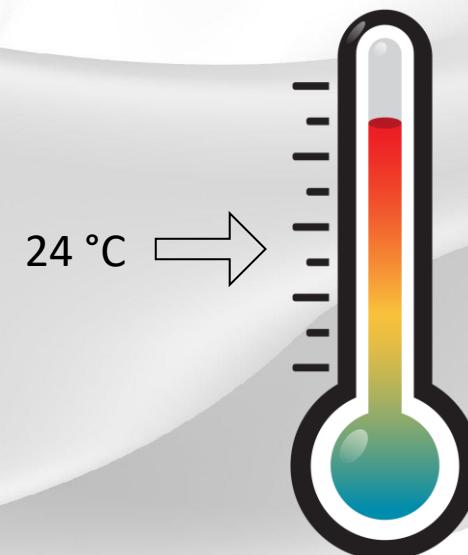
Clasificación

¿Cómo estará el clima mañana, hará frío o calor?

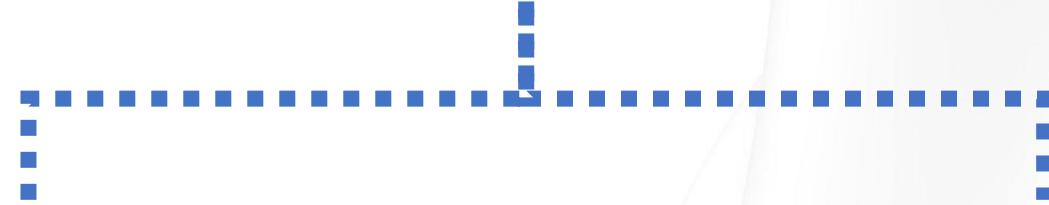


Regresión

¿Qué temperatura hará mañana?



Aprendizaje Supervisado



Clasificación

Churn ¿El cliente se fue?



Estimamos con base a los clientes que se fueron en meses anteriores

Regresión

Estimación ingreso próximo trimestre



Estimamos con base a los ingresos de trimestres anteriores

Aprendizaje Supervisado

Clasificación

Churn ¿El cliente se fue?



Estimamos con base a los clientes que se fueron en meses anteriores

Regresión

Estimación ingreso próximo trimestre

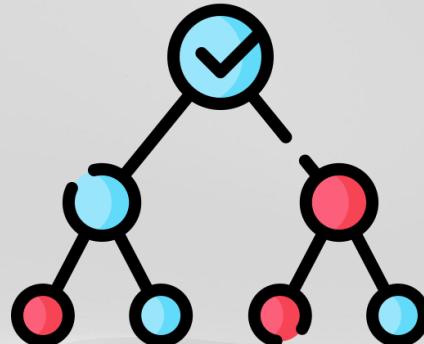


Estimamos con base a los ingresos de trimestres anteriores

Aprendizaje Supervisado

Clasificación

Arboles de Decisión

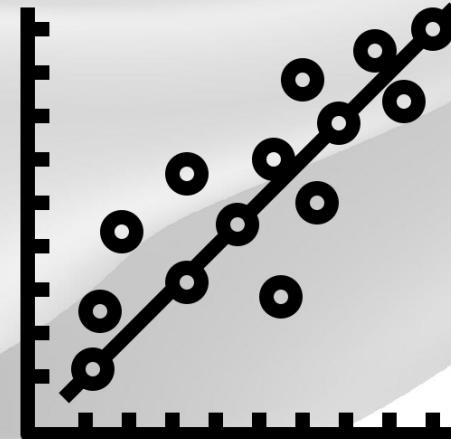


Regresión logística



Regresión

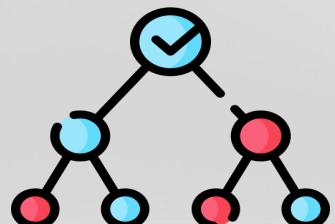
Regresión lineal



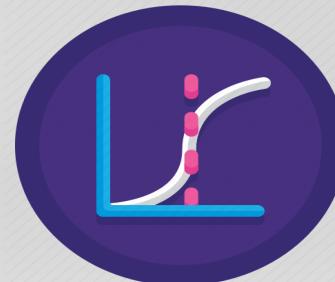
Aprendizaje Supervisado

Clasificación

Arboles de
Decisión

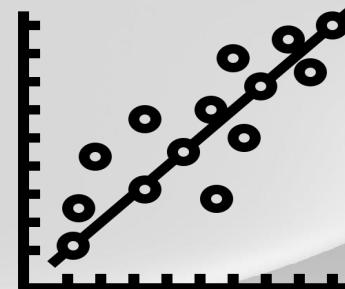


Regresión
logística



Regresión

Regresión
lineal



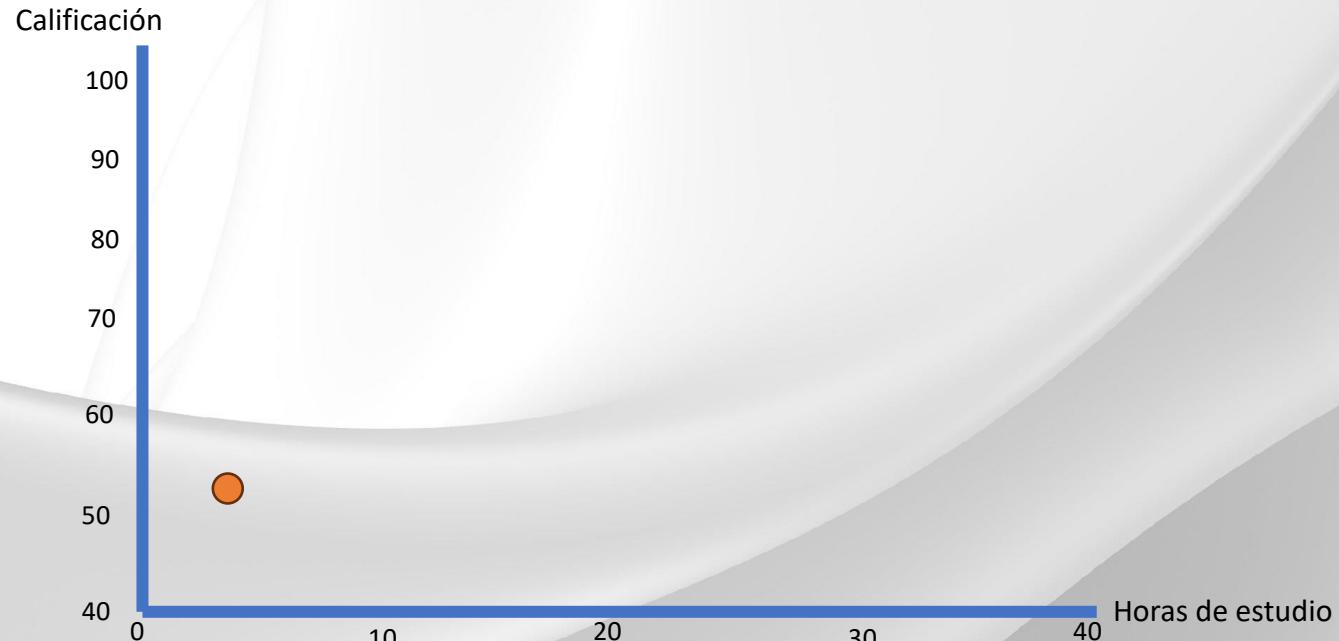
Estos son **algunos** de los
muchos modelos que
existen



Modelos Regresión: Regresión Lineal

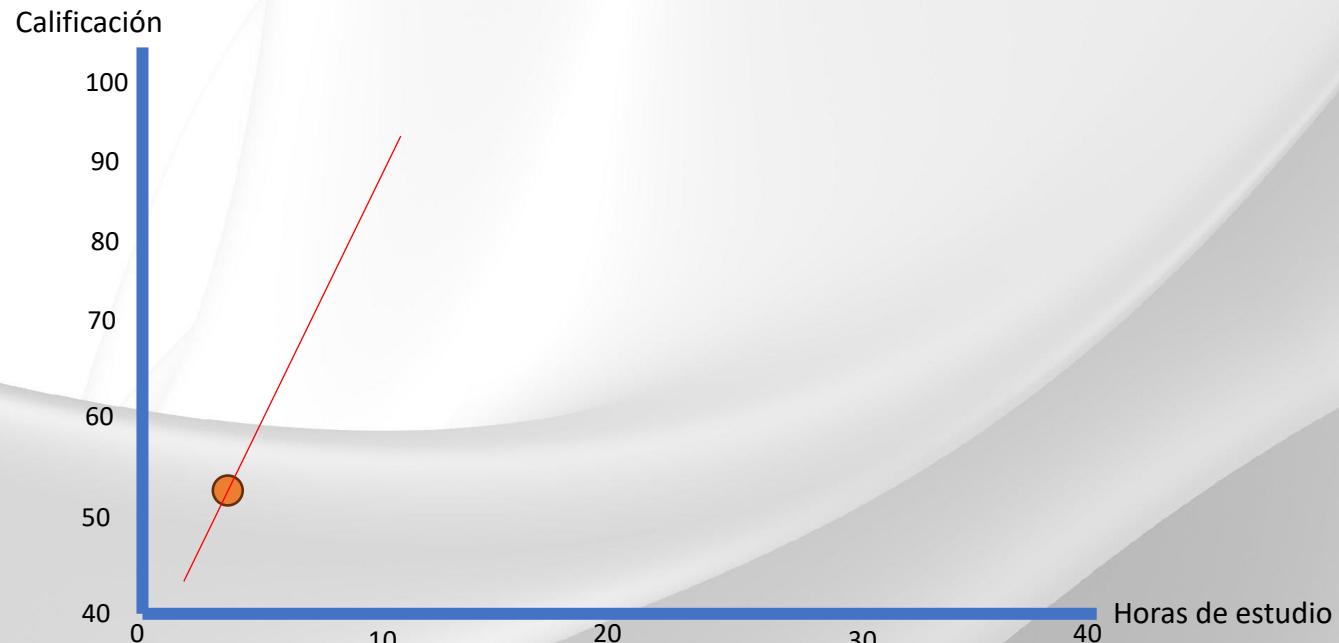
Regresión Lineal

- Un modelo de regresión lineal es como trazar una línea recta a través de un grupo de puntos en un gráfico
- Lo que hace la regresión lineal es encontrar la línea recta que mejor se ajusta a todos esos puntos de datos.
- Una vez que tienes esa línea, puedes usarla para predecir



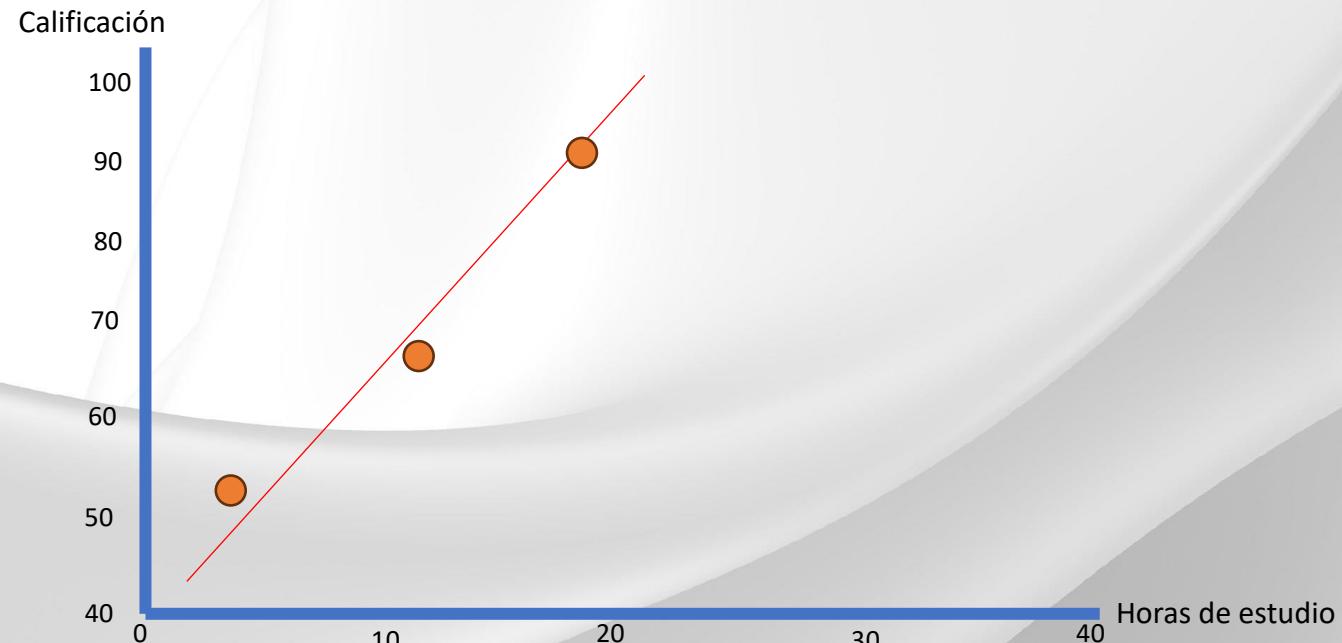
Regresión Lineal

- Un modelo de regresión lineal es como trazar una línea recta a través de un grupo de puntos en un gráfico
- Lo que hace la regresión lineal es encontrar la línea recta que mejor se ajusta a todos esos puntos de datos.
- Una vez que tienes esa línea, puedes usarla para predecir



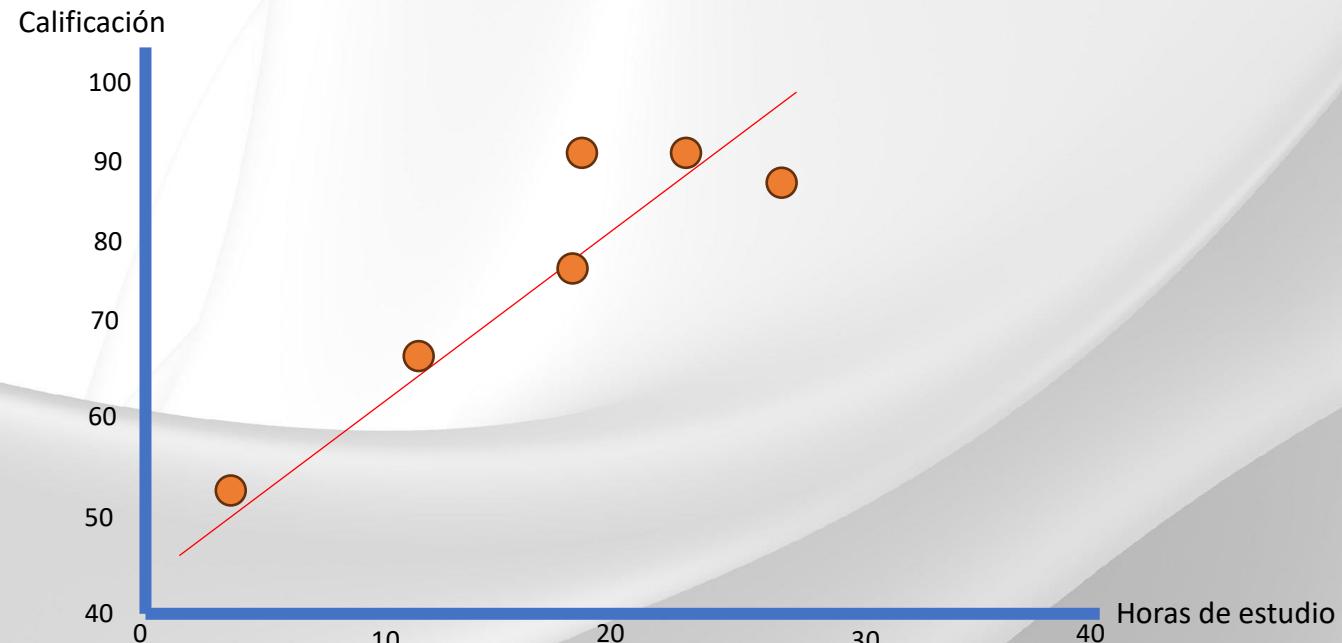
Regresión Lineal

- Un modelo de regresión lineal es como trazar una línea recta a través de un grupo de puntos en un gráfico
- Lo que hace la regresión lineal es encontrar la línea recta que mejor se ajusta a todos esos puntos de datos.
- Una vez que tienes esa línea, puedes usarla para predecir



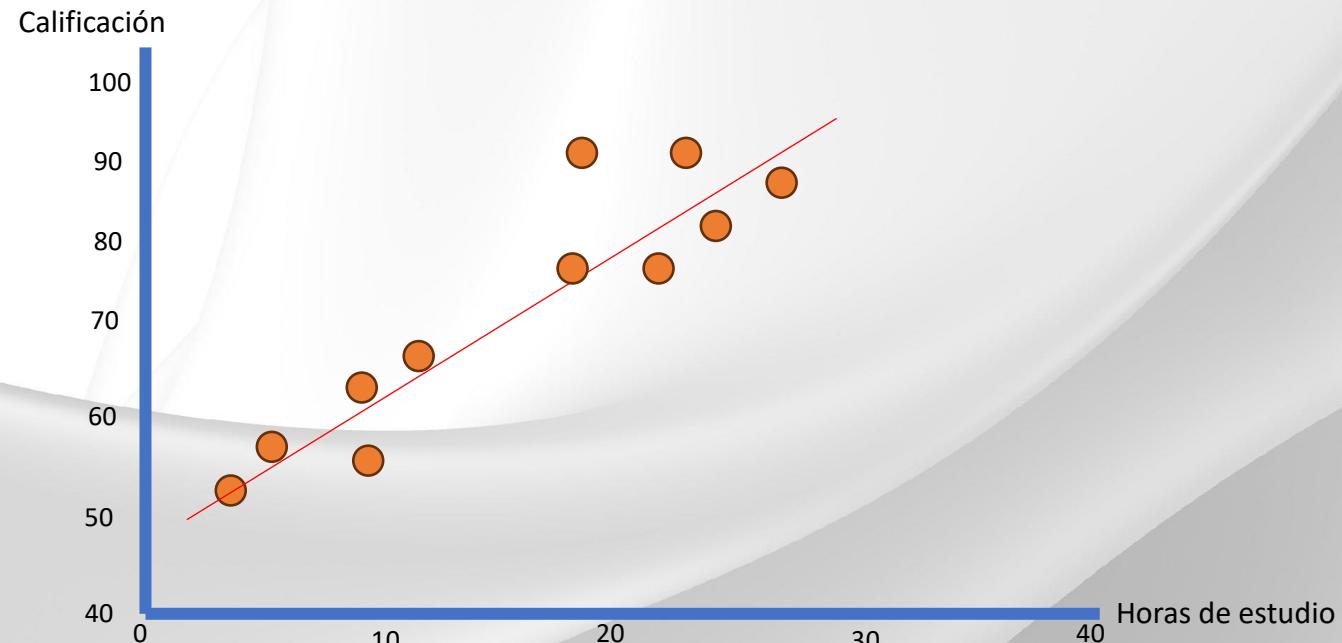
Regresión Lineal

- Un modelo de regresión lineal es como trazar una línea recta a través de un grupo de puntos en un gráfico
- Lo que hace la regresión lineal es encontrar la línea recta que mejor se ajusta a todos esos puntos de datos.
- Una vez que tienes esa línea, puedes usarla para predecir



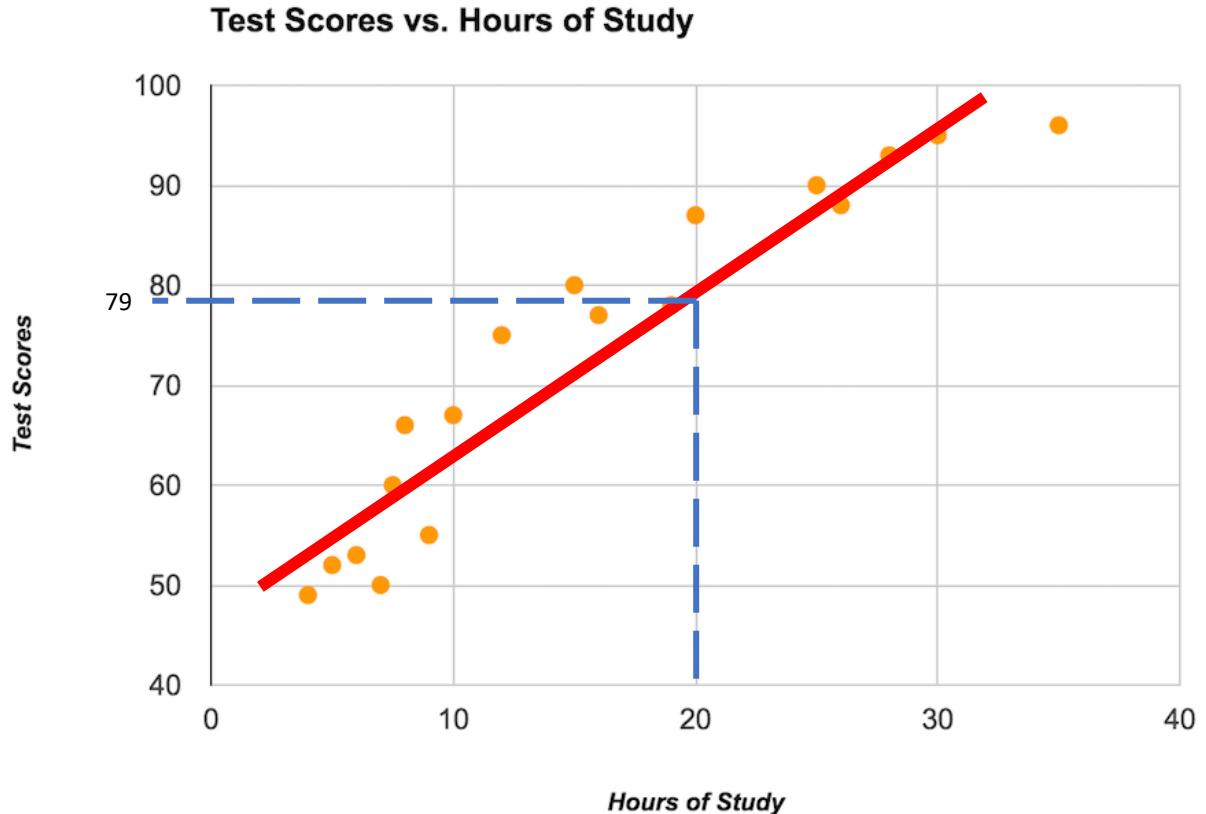
Regresión Lineal

- Un modelo de regresión lineal es como trazar una línea recta a través de un grupo de puntos en un gráfico
- Lo que hace la regresión lineal es encontrar la línea recta que mejor se ajusta a todos esos puntos de datos.
- Una vez que tienes esa línea, puedes usarla para predecir



Regresión Lineal

- Supongamos que quieres predecir la calificación que alguien obtendrá en un examen basándote en cuántas horas ha estudiado
- La predicción cuando se estudia 20 horas es una calificación aproximada de 79
- La principal ventaja de la regresión lineal es que es una manera simple y efectiva de ver y entender relaciones entre dos cosas, y hacer predicciones basadas en esos datos.

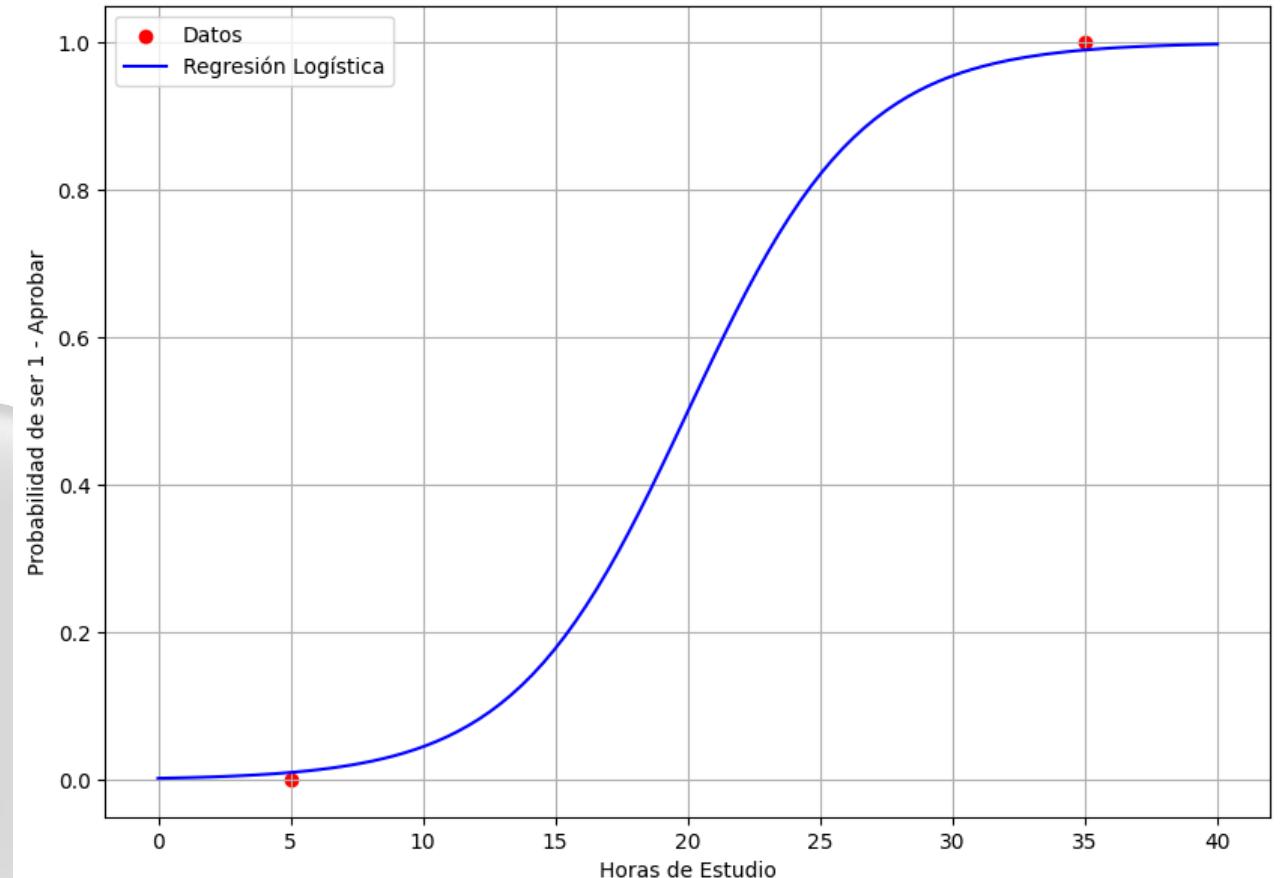




Modelos Clasificación: Regresión Logistica

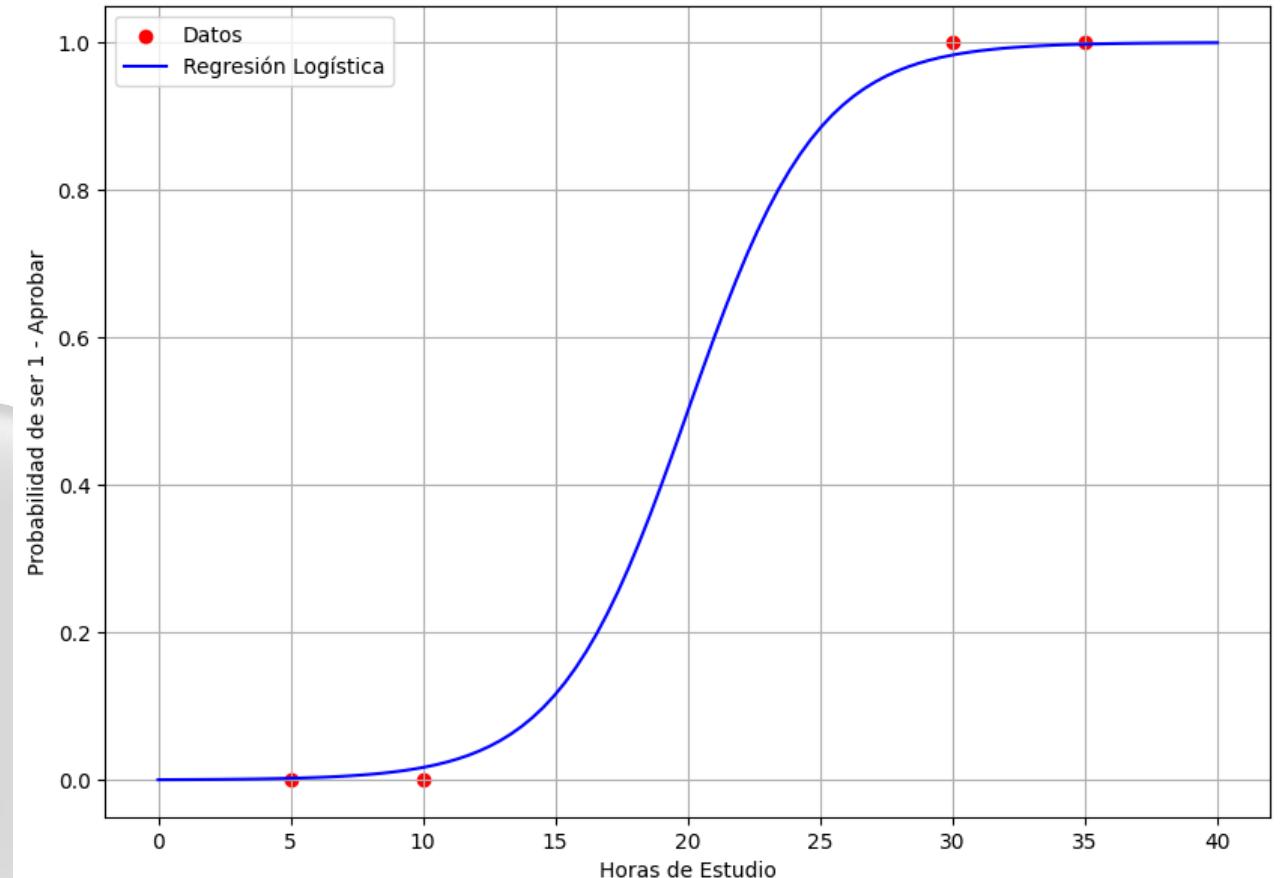
Regresión Logística

- Un modelo de regresión logística es como un sistema de toma de decisiones que se enfrenta a preguntas de "sí"(1) o "no"(0)
- El modelo toma diferentes factores en cuenta y calcula la probabilidad de que la respuesta sea "sí".
- En lugar de darte una respuesta directa, te da un porcentaje o probabilidad.



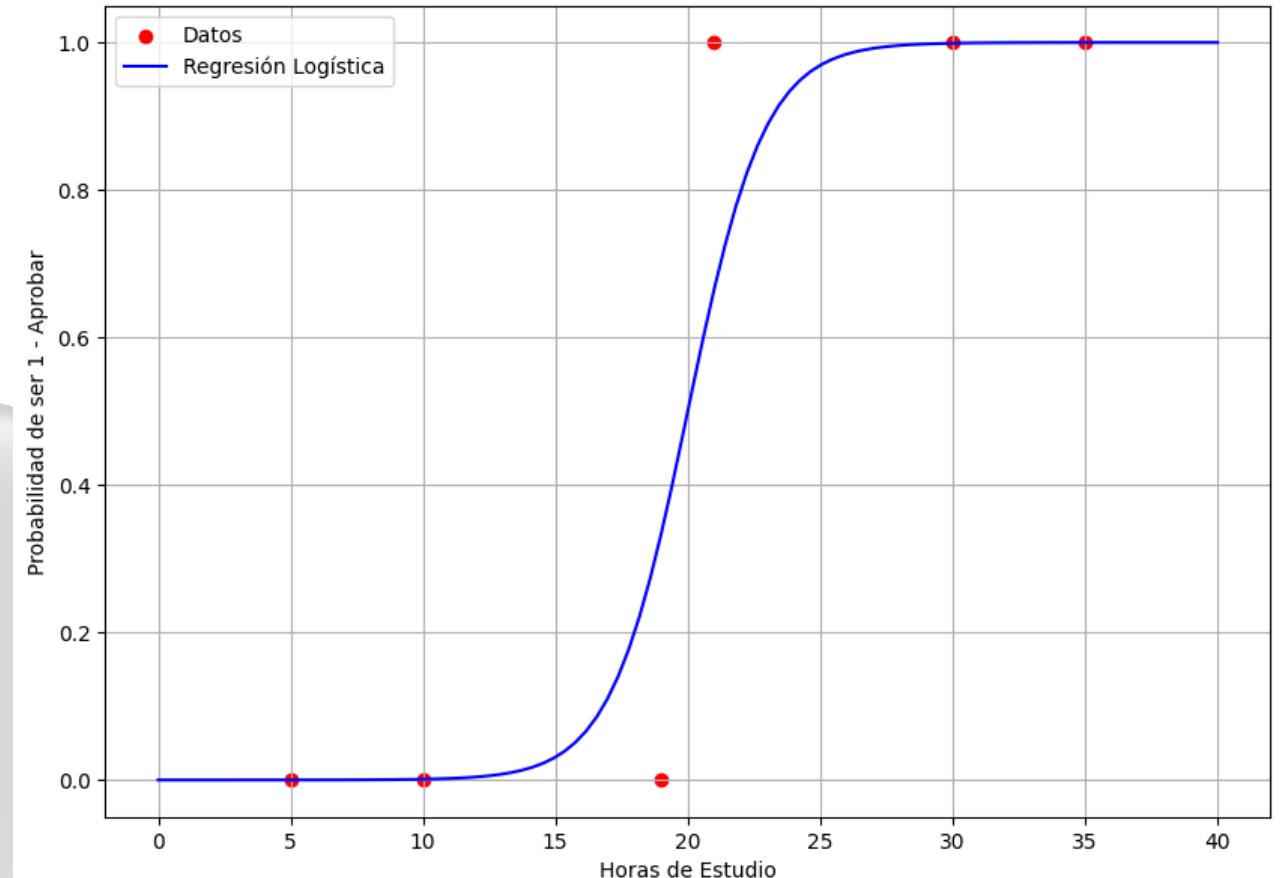
Regresión Logística

- Un modelo de regresión logística es como un sistema de toma de decisiones que se enfrenta a preguntas de "sí"(1) o "no"(0)
- El modelo toma diferentes factores en cuenta y calcula la probabilidad de que la respuesta sea "sí".
- En lugar de darte una respuesta directa, te da un porcentaje o probabilidad.



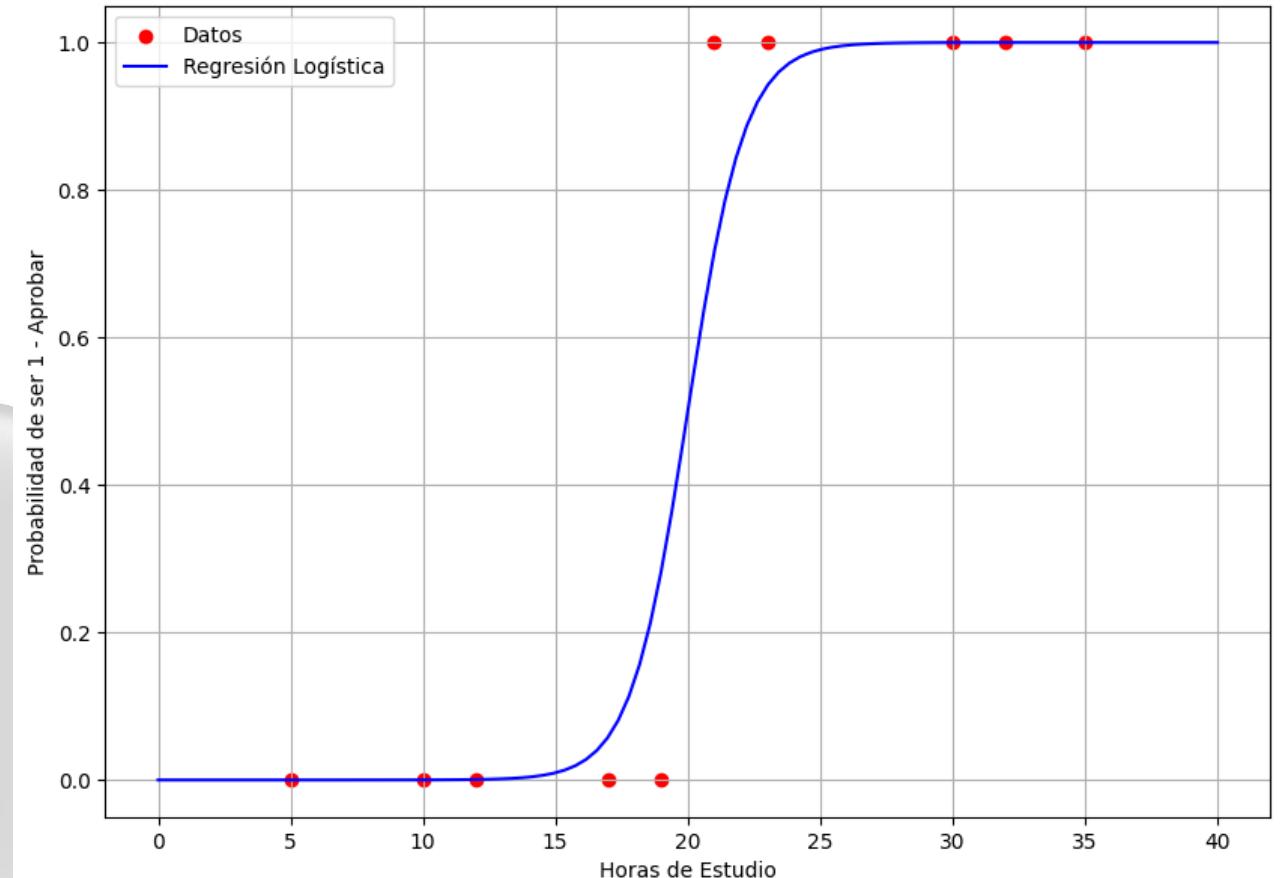
Regresión Logística

- Un modelo de regresión logística es como un sistema de toma de decisiones que se enfrenta a preguntas de "sí"(1) o "no"(0)
- El modelo toma diferentes factores en cuenta y calcula la probabilidad de que la respuesta sea "sí".
- En lugar de darte una respuesta directa, te da un porcentaje o probabilidad.



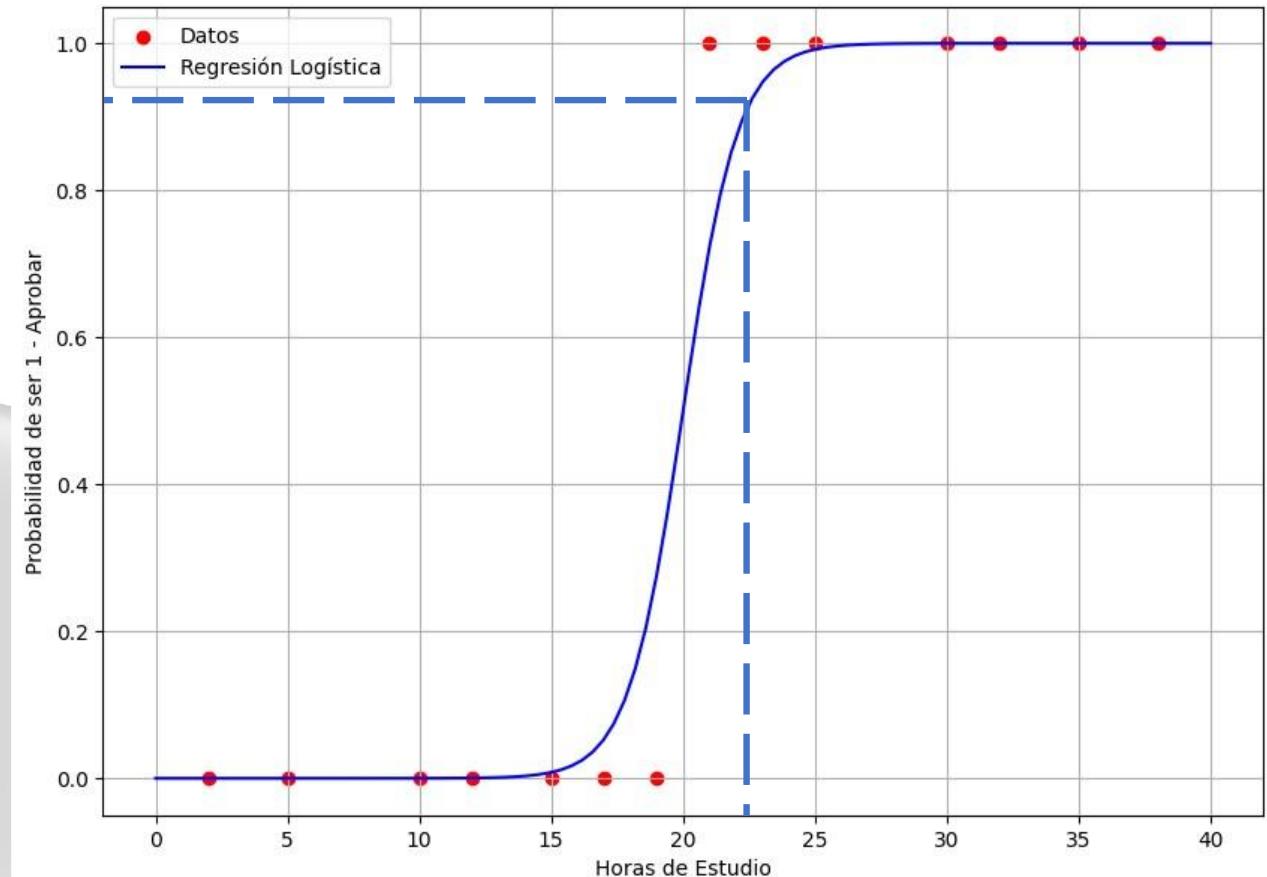
Regresión Logística

- Un modelo de regresión logística es como un sistema de toma de decisiones que se enfrenta a preguntas de "sí"(1) o "no"(0)
- El modelo toma diferentes factores en cuenta y calcula la probabilidad de que la respuesta sea "sí".
- En lugar de darte una respuesta directa, te da un porcentaje o probabilidad.



Regresión Logística

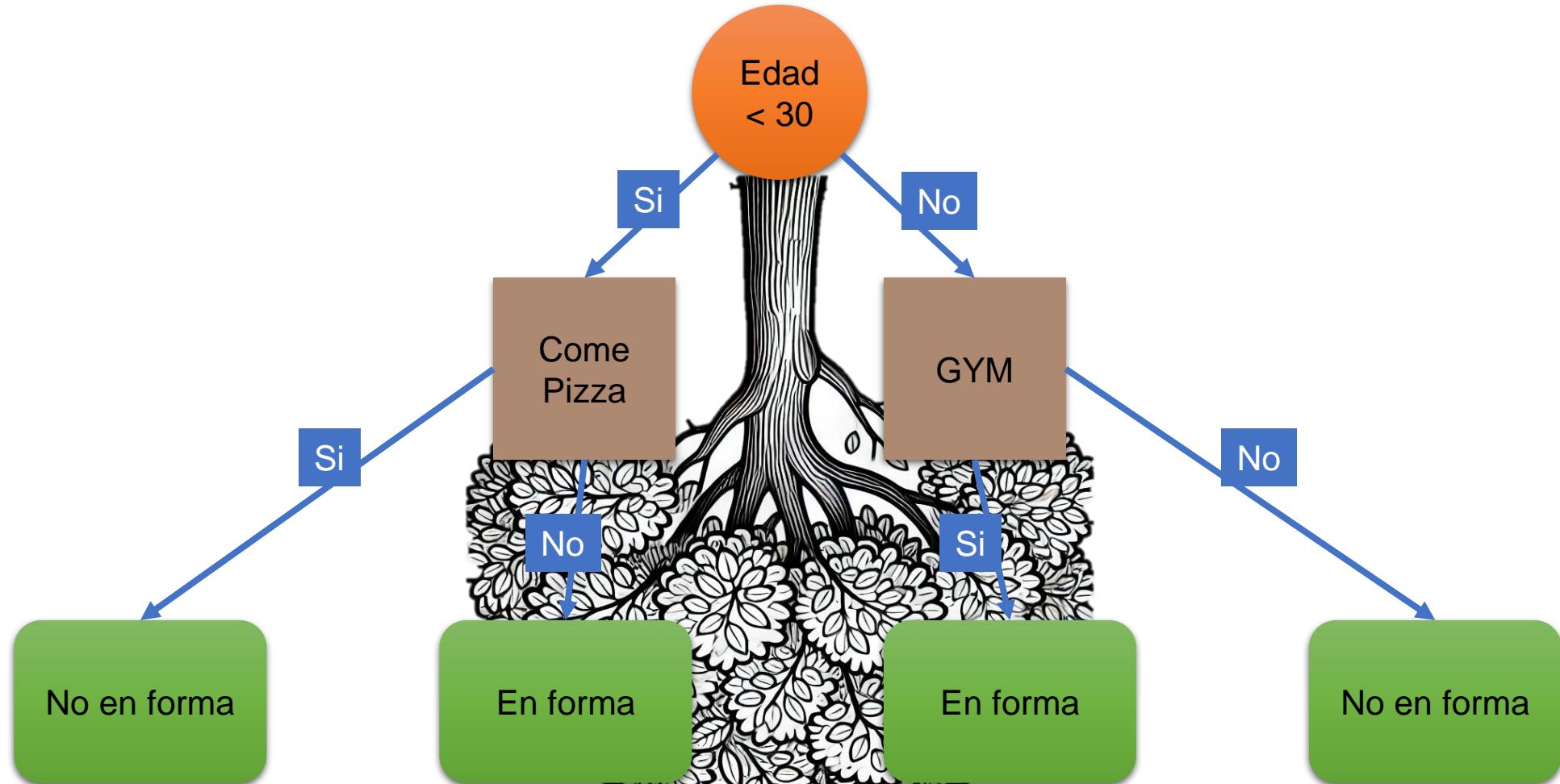
- Supongamos que quieres predecir si alguien aprobará (1) en un examen basándote en cuántas horas ha estudiado
- La predicción cuando se estudia 23 horas es que la persona aprobará, debido a que, la probabilidad es aproximadamente 93%



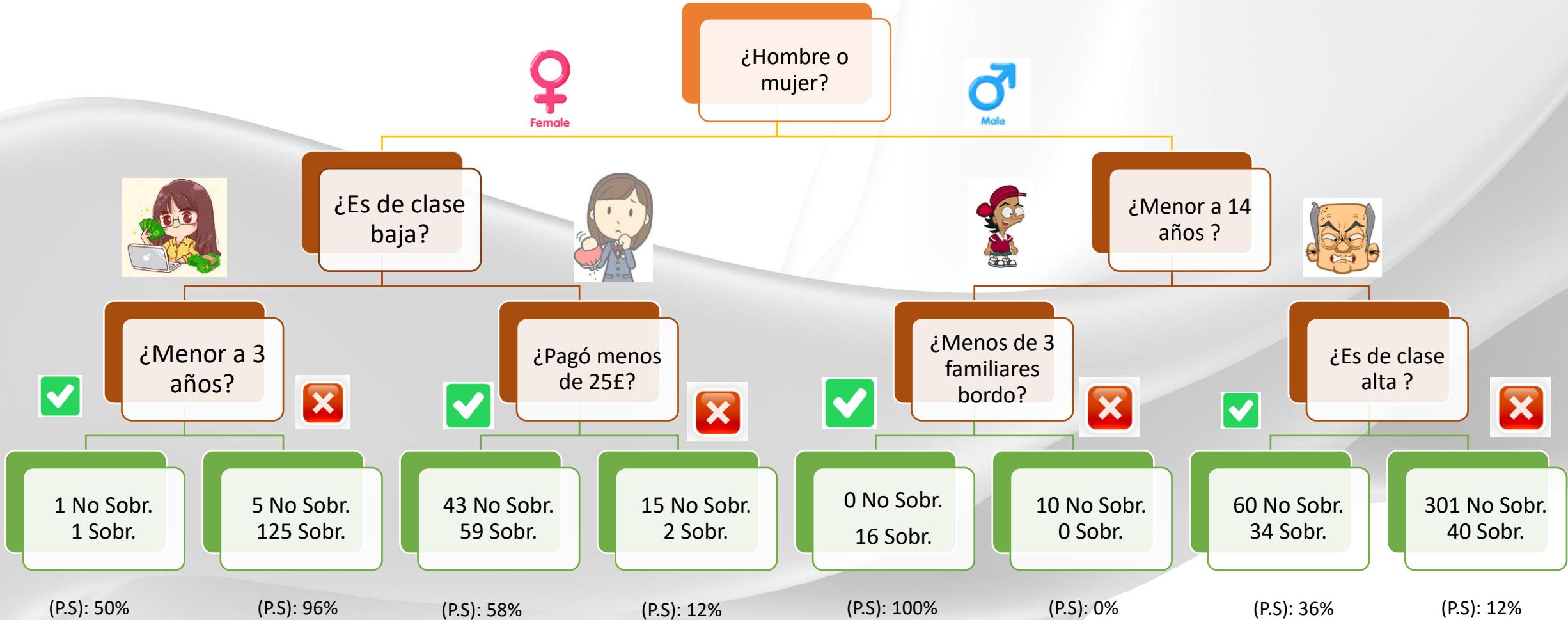


Modelos Clasificación: Árboles

	nodo raíz
	nodos internos
	nodos hojas
	ramas



Prob. Supervivencia (P.S): 38%





Ejercicio

¿Qué modelo utilizaría en cada caso?

Caso 1:

Predicción de Churn de Clientes

Predecir si un cliente dejará de utilizar los servicios de la empresa (churn) basado en variables como la duración de la relación, el uso mensual, el número de quejas, etc.



Caso 2:

Determinación de la Calidad del Servicio

Predecir si un cliente reportará problemas con la calidad del servicio en función de la región, la cobertura de red, y la antigüedad del cliente.



Caso 3:

Predicción del Ingreso Mensual de Prepago

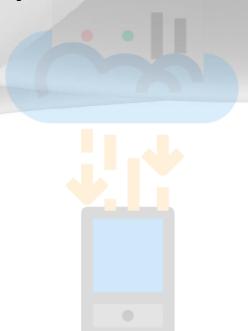
Estimar el ingreso mensual que generará el segmento prepago en el próximo trimestre, tomando en cuenta su ingreso histórico.



Caso 4:

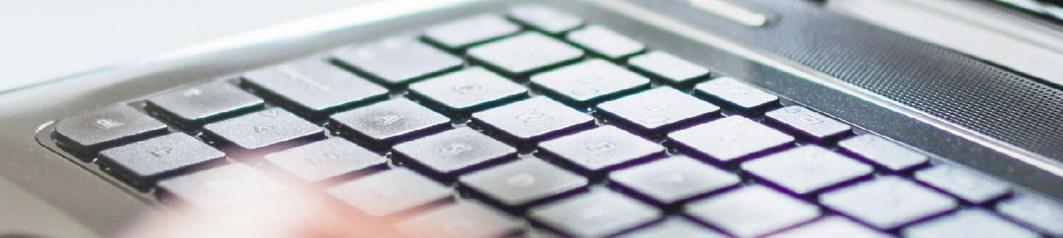
Estimación del Consumo Mensual de Datos

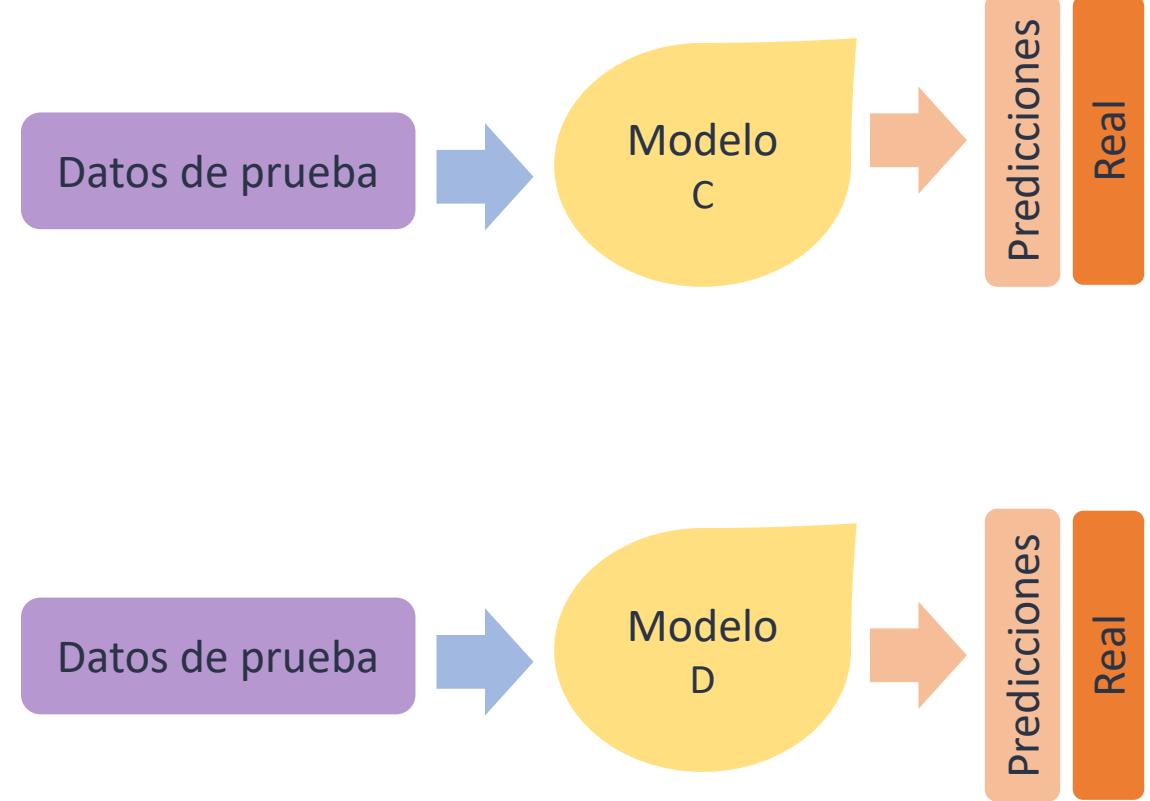
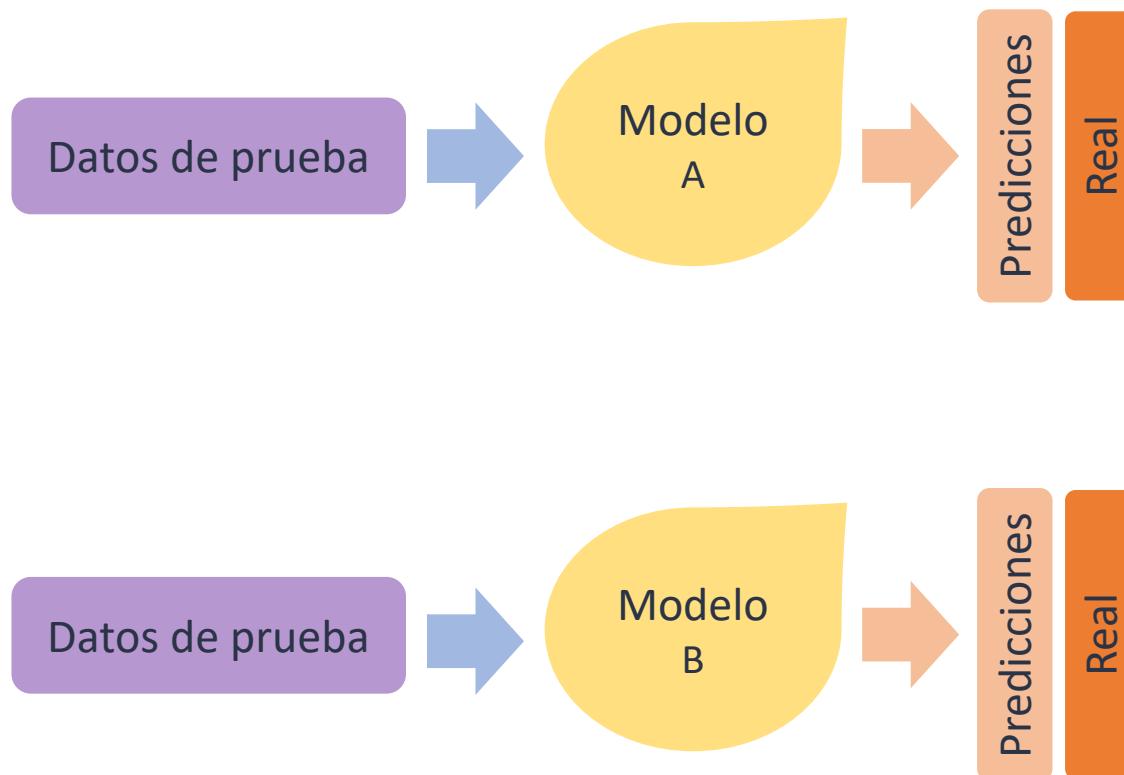
Predecir la cantidad de datos móviles que un cliente consumirá el próximo mes basado en su historial de consumo, tipo de plan, y otros factores.





Métricas de Desempeño

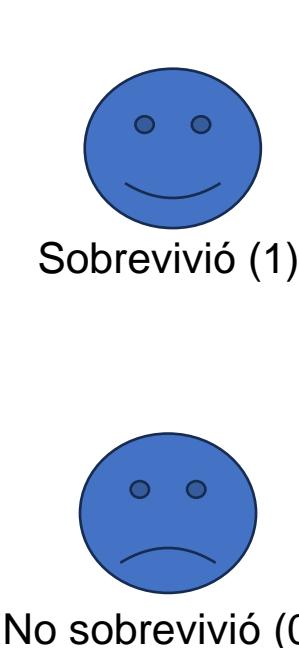




¿Cuál elegimos?

Métricas de desempeño

Matriz de Confusión



		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	El modelo predijo 0 y el real es 0	El modelo predijo 1 y el real es 0
	Sobrevivió (1)	El modelo predijo 0 y el real era 1	El modelo predijo 1 y el real era 1

La matriz de confusión es una herramienta que se utiliza en clasificación para evaluar el rendimiento de un modelo

Métricas de desempeño

Matriz de Confusión

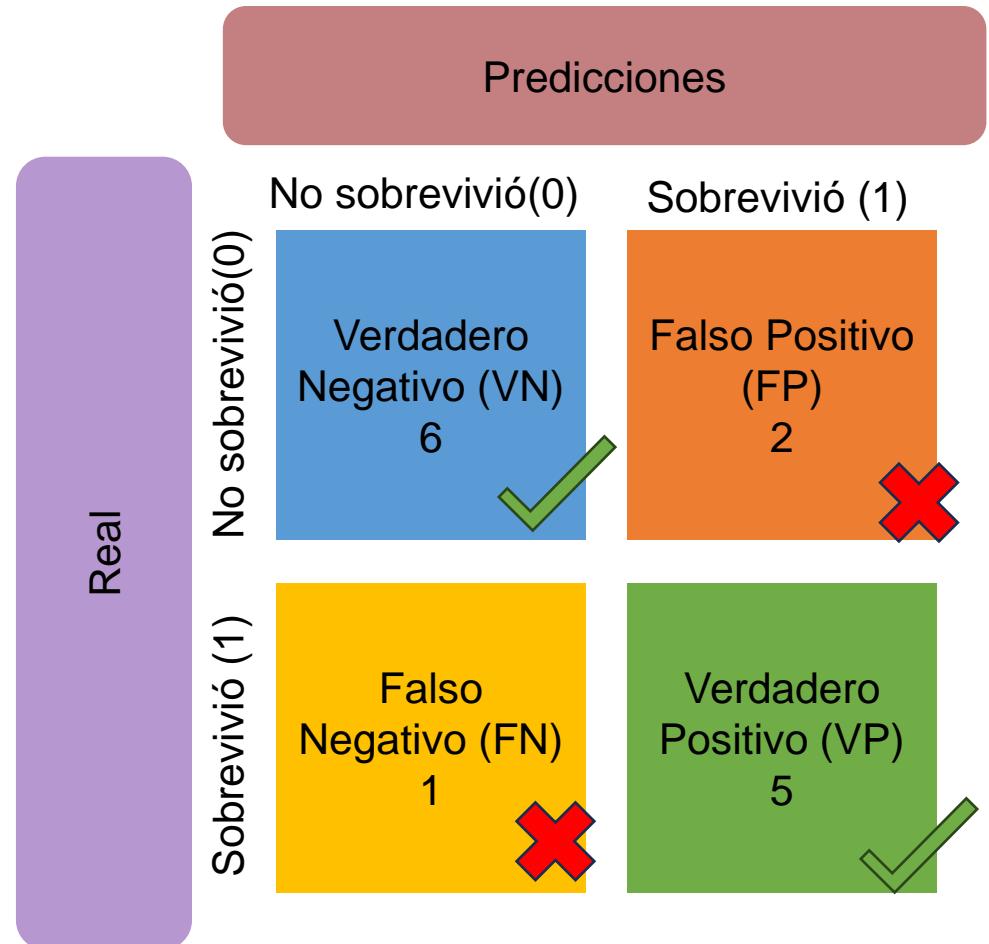
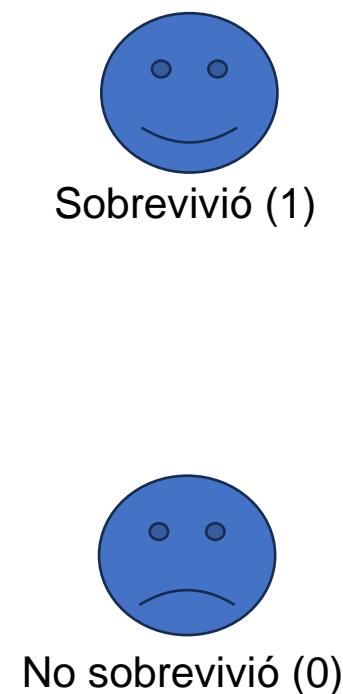
		Predicciones				
		No sobrevivió(0)	Sobrevivió (1)			
Real	No sobrevivió(0)	El modelo predijo 0 y el real es 0	El modelo predijo 1 y el real es 0	Real	No sobrevivió(0)	
	Sobrevivió (1)	El modelo predijo 0 y el real era 1	El modelo predijo 1 y el real era 1			
		 		 		
		 		 		

14 pasajeros

Real	Predicción	Tipo
0	0	VN
1	1	VP
0	0	VN
0	0	VN
1	0	FN
0	0	VN
0	1	FP
0	1	FP
1	1	VP
1	1	VP
0	0	VN
0	0	VN
1	1	VP
1	1	VP

Métricas de desempeño

Matriz de Confusión

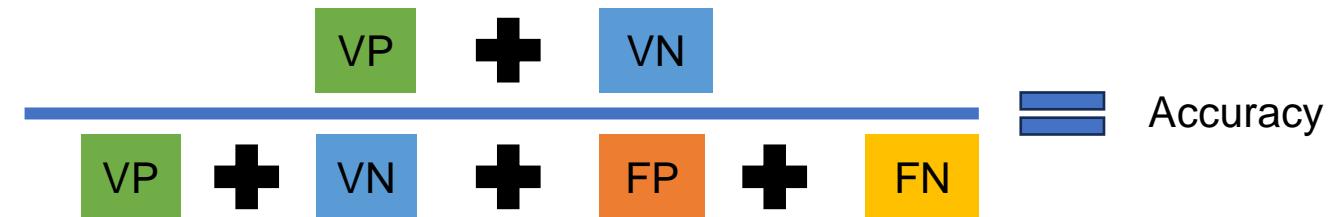


Métricas de desempeño

		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	Verdadero Negativo (VN)	Falso Positivo (FP)
	Sobrevivió (1)	Falso Negativo (FN)	Verdadero Positivo (VP)

Accuracy – Exactitud

En general, ¿Con qué frecuencia son correctas las predicciones?

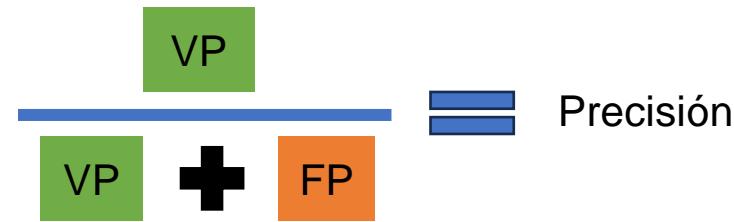


Métricas de desempeño

		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	Verdadero Negativo (VN)	Falso Positivo (FP)
	Sobrevivió (1)	Falso Negativo (FN)	Verdadero Positivo (VP)

Precisión

Cuando predice 1, ¿Con qué frecuencia es correcto?



¿De todas las veces que el modelo predice que un pasajero sobrevivirá, cuántos realmente sobreviven?

Métricas de desempeño

Precisión

¿Cuándo utilizarlo?

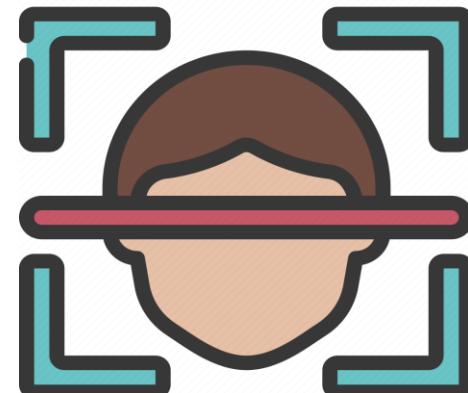
Cuando es importante que las predicciones positivas (1) sean lo más correctas posible, minimizando los falsos positivos.

Útil en situaciones donde el costo de un falso positivo es alto.

Clasificación de spam: Es más importante que los correos clasificados como spam realmente lo sean, para evitar que correos importantes se pierdan.



Sistemas que detectan rostros: como los usados para desbloquear teléfonos o en seguridad, es crucial que el rostro identificado sea el correcto. Un falso positivo podría permitir el acceso a alguien no autorizado.

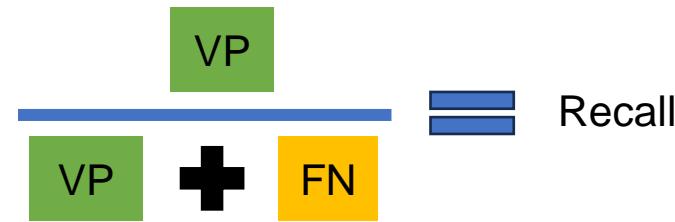


Métricas de desempeño

		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	Verdadero Negativo (VN)	Falso Positivo (FP)
	Sobrevivió (1)	Falso Negativo (FN)	Verdadero Positivo (VP)

Recall - Sensibilidad

Cuando en realidad es un 1, ¿Con qué frecuencia predice un 1?



De todos los pasajeros que realmente sobrevivieron, ¿cuántos logra identificar correctamente mi modelo?

Métricas de desempeño

Recall

¿Cuándo utilizarlo?

Cuando es importante capturar todos los casos positivos (1), incluso si se corre el riesgo de clasificar algunos negativos como positivos.

Útil en situaciones donde perder un caso positivo es costoso o peligroso.

Detección de enfermedades: Es crucial identificar a todos los pacientes enfermos, incluso si algunos sanos se identifican erróneamente como enfermos.

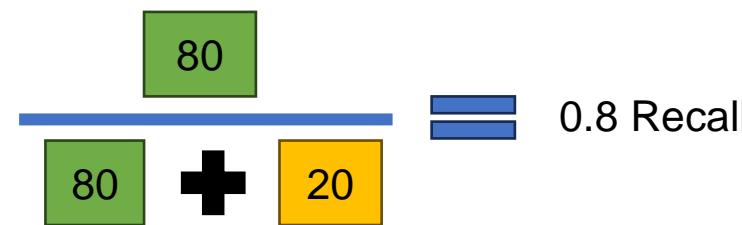
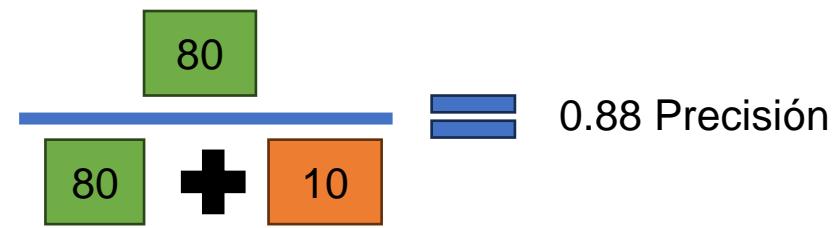


Detección de fraudes: Capturar todos los posibles fraudes es prioritario, aunque se etiqueten algunas transacciones legítimas como fraudulentas.



Métricas de desempeño

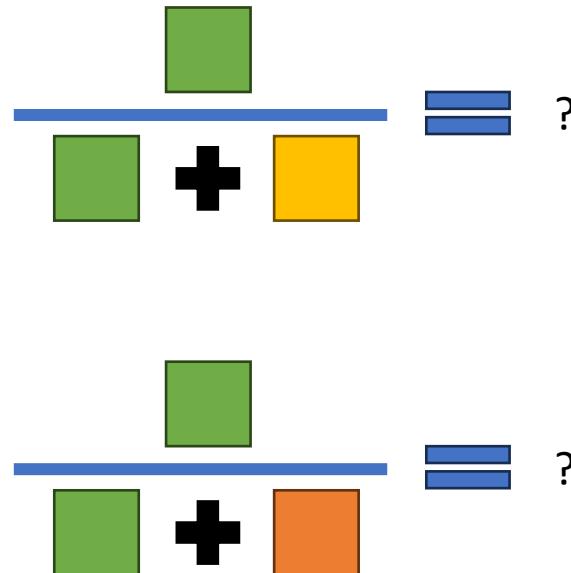
		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	Verdadero Negativo (VN) 10	Falso Positivo (FP) 20
	Sobrevivió (1)	Falso Negativo (FN) 10	Verdadero Positivo (VP) 80



Métricas de desempeño



		Predicciones	
		No sobrevivió(0)	Sobrevivió (1)
Real	No sobrevivió(0)	669	74
	Sobrevivió (1)	57	836





Analítica avanzada