Inteligencia Artificial

Instituto Nacional de Astrofísica, Óptica y Electrónica

Presents:

MSc. Mireya Lucia Hernández Jaimes

Content

1. Quizz
2. Conceptos fundamentales
3. Análisis de los conjuntos de datos
4. Pre-procesamiento de los datos
5. Aprendizaje de máquina

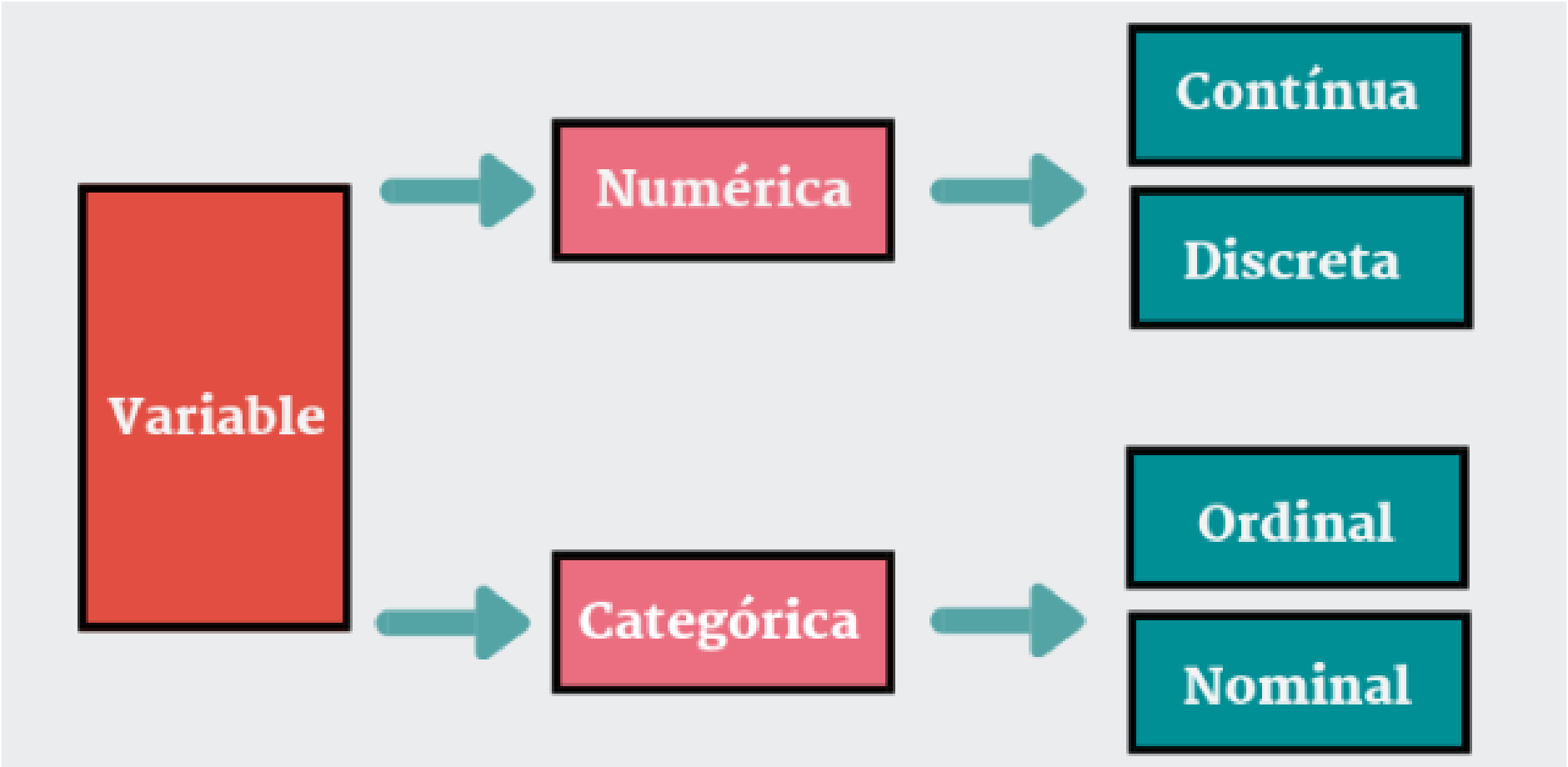
Quizz

- Menciona 3 aplicaciones de IA
- Define IA
- Define ML
- Define DL
- ¿Cuál es la relacion entre IA-ML-DL?
- ¿Cuáles son las similitudes y diferencias entre IA-ML-DL?
- ¿Cuáles son las diferencias entre IA-ML-DL?
- ¿Cómo representamos las propiedades/características/atributos de los objetos/instancias?

En la presentación anterior se discutieron los siguientes conceptos:

- Instancia/muestra/ejemplo
- Vectores de características "Feature Vectors"

Tipos de datos/variables



Clasifica los siguientes tipos de datos/variables

- Nivel de ansiedad
- Edad
- Grupo sanguineo
- Número de pacientes al día
- Cantidad de glucosa en la sangre
- Sexo: F/M
- Evaluación de un servicio (Malo, Regular, Bueno, Muy Bueno, Excelente)
- Nivel de estudio
- País de origen
- Tamaño de ropa (XS, S, M, L, XL)

Conjunto de datos "Datasets" Es una colección de datos, atributos, características y especificaciones.

No.	Number of times pregnant	Plasma glucose concentration	Diastolic blood pressure	Triceps skin fold thickness	2-Hour serum insulin	Body mass index	Diabetes pedigree function	Age	Diabetes
1	6	148	72	35	0	33.6	0.627	50	tested_positive
2	1	85	66	29	0	26.6	0.351	31	tested_negative
3	8	183	64	0	0	23.3	0.672	32	tested_positive
4	1	89	66	23	94	28.1	0.167	21	tested_negative
5	0	137	40	35	168	43.1	2.288	33	tested_positive
6	5	116	74	0	0	25.6	0.201	30	tested_negative
7	3	78	50	32	88	31.0	0.248	26	tested_positive
8	10	115	0	0	0	35.3	0.134	29	tested_negative
9	2	197	70	45	543	30.5	0.158	53	tested_positive
10	8	125	96	0	0	0.0	0.232	54	tested_positive
11	4	110	92	0	0	37.6	0.191	30	tested_negative
12	10	168	74	0	0	38.0	0.537	34	tested_positive
13	10	139	80	0	0	27.1	1.441	57	tested_negative
14	1	189	60	23	846	30.1	0.398	59	tested_positive

	outlook	temp	humidity	windy	play
0	sunny	hot	high	False	no
1	sunny	hot	high	True	no
2	overcast	hot	high	False	yes
3	rainy	mild	high	False	yes
4	rainy	cool	normal	False	yes
5	rainy	cool	normal	True	no

- Tamaño del dataset: Número de instancias, Número de atributos
- Identificar tipos de datos (numéricos, categóricos)
- Identificar las etiquetas (clases)
- Analizar las distribuciones de las clases

Pre-procesamiento: se refiere a las técnicas de preparación (limpieza y organización) de los datos sin procesar ("Raw Data") para que sean adecuados para los modelos de ML.

- Identificar los valores faltantes.
- Identificar los valores duplicados.
- Codificar variables categóricas.
- Normalizar los datos numéricos.
- Estandarizar los datos numéricos.
- Ingeniería de Características (Feature engineering: Selección y extracción de características).
- Balancear el conjunto de datos.

Codificar Variables categóricas: Transformar las variables categóricas a variables numéricas.

- Label encoding
- One-Hot encoding

Label encoding: Es un proceso adecuado para datos categóricos ordinales. Es decir, tienen un orden claro.

Height	Height
Tall	0
Medium	1
Short	2

One-Hot encoding: s un proceso adecuado para datos categóricos nominales. Asigna a qué clase pertenece la instancia con un 1.

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

Normalización de los datos: reescala las características individuales para que se ajusten a un rango específico, normalmente [0, 1].

-
- Min-max

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Min-Max

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	[redacted]	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		Range 0-1

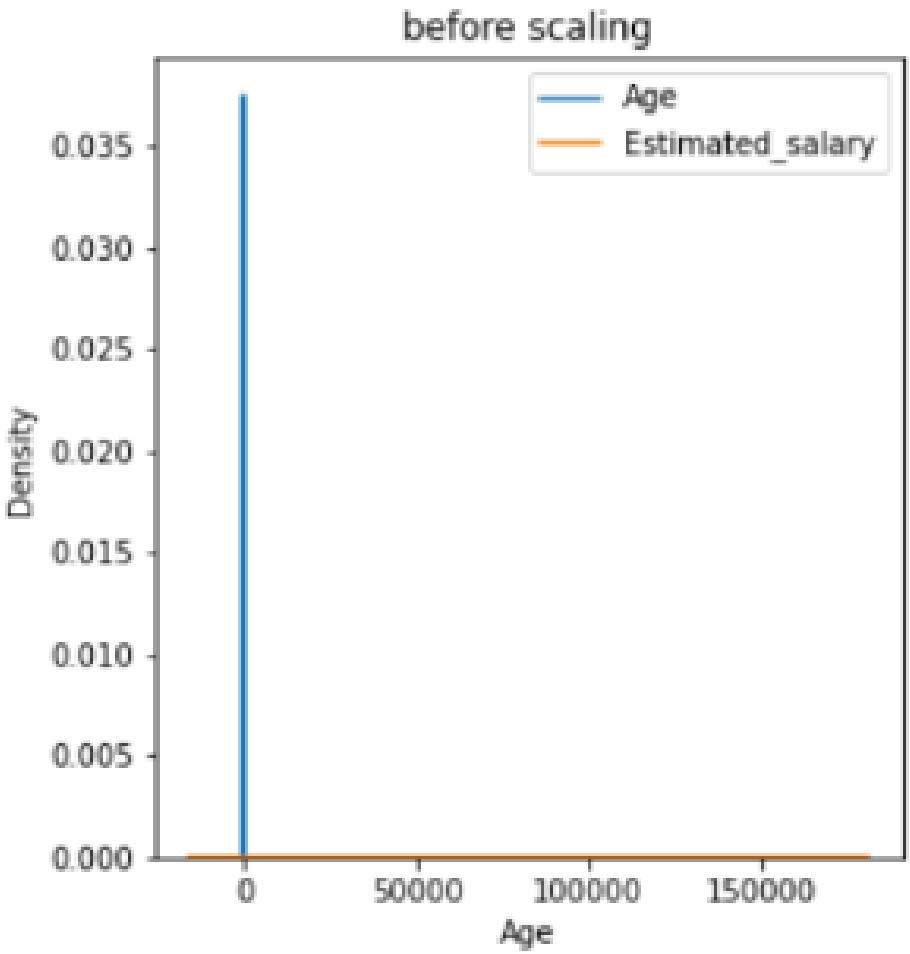
Min-Max

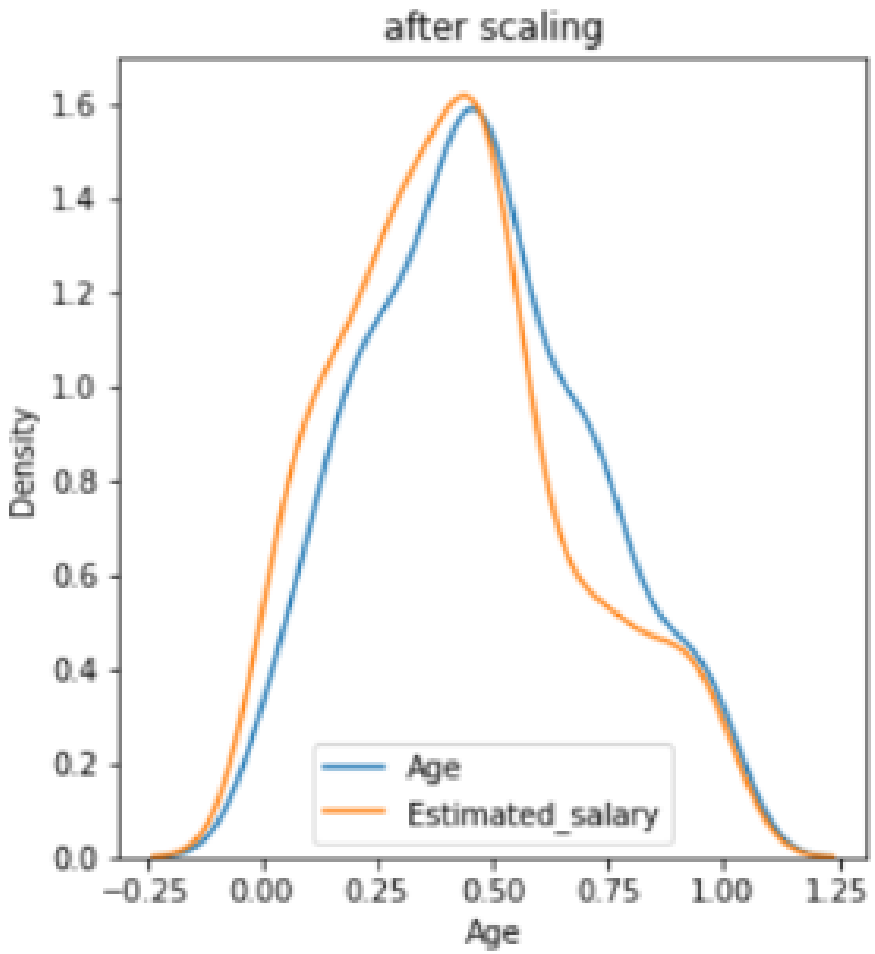
#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		Range 0-1

Min-Max


before scaling
Age
Estimated_salary
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
50000
100000
150000
Age


after scaling
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
Age
Estimated_salary
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Age

Estandarización de los datos: transforma los datos para que tengan una media de 0 y una desviación estándar de 1.

- Z-score

$$\text{New value} = (x - \mu) / \sigma$$

Z-score

Data
3
5
5
8
9
12
12
13
15
16
17
19
22
24
25
134

De estos datos se tiene una media de 21.2 y una desviación estandar de 29.8

Z-score

Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	[redacted]
12	-0.31
12	-0.31
13	-0.28

Z-score

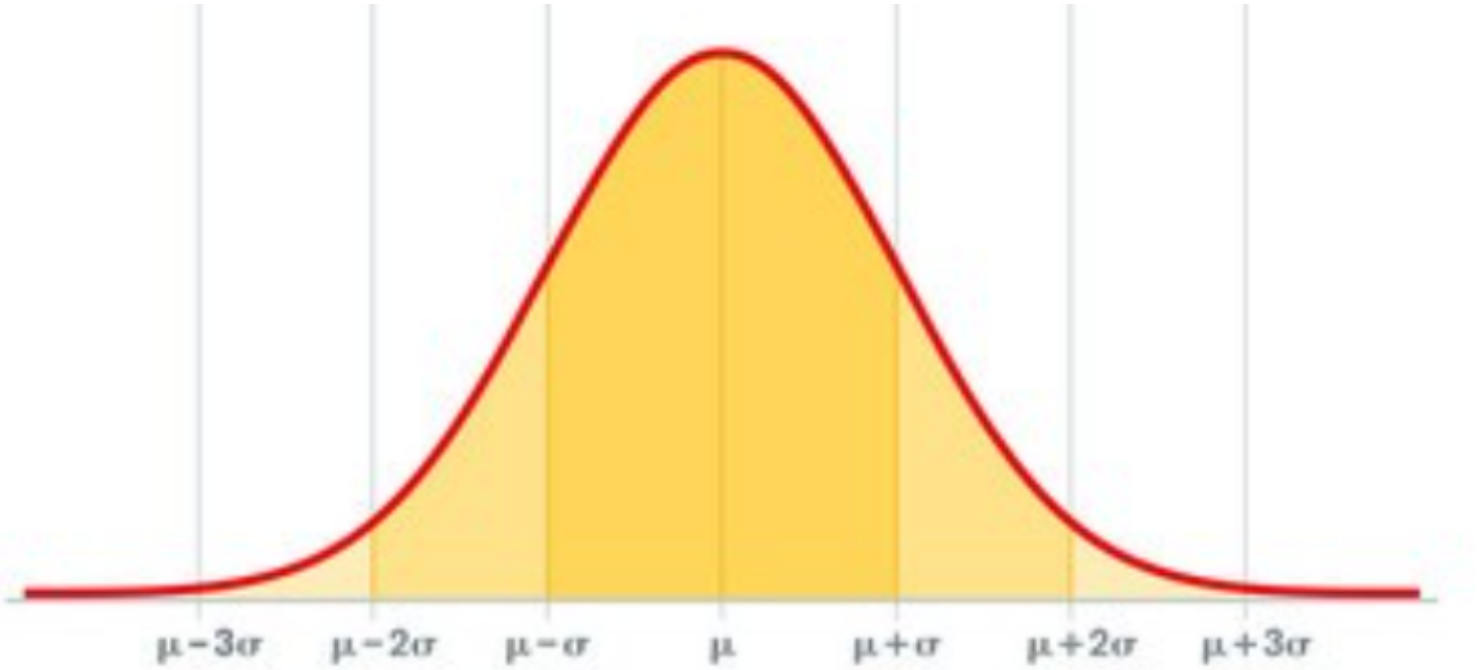
Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28

Z-score

Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28

Z-score

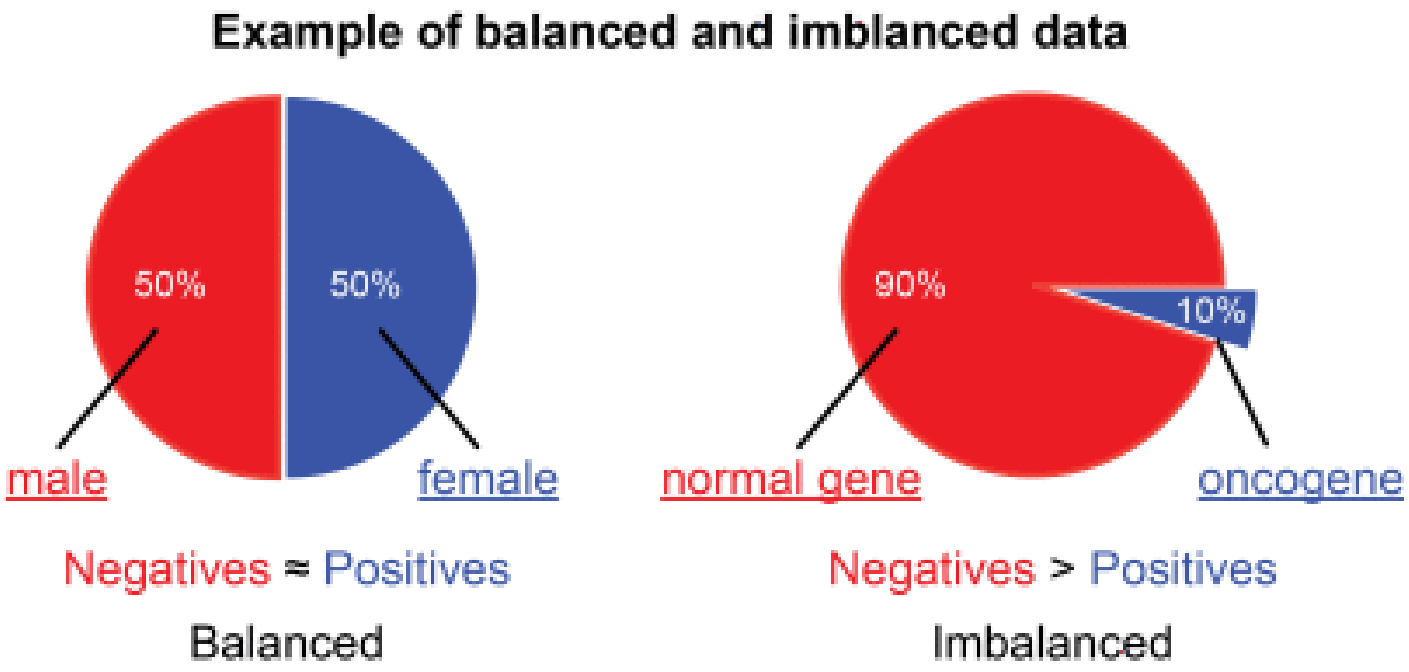
¿Qué me representa el z-score?



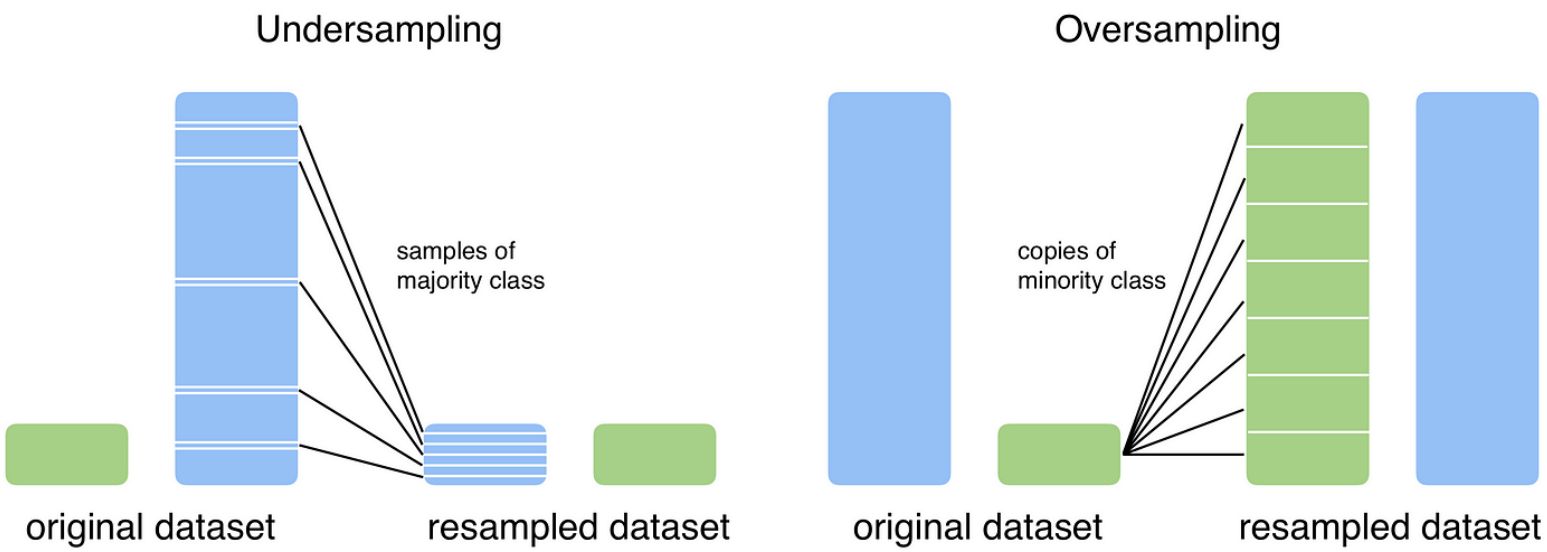
Feature engineering:

- Extracción de características: Extraer información significativa de los datos para crear nuevas características de las existentes.
- Selección de características:
  - Reducir la dimensionalidad de los datos.
  - Selección de caracerísticas relevantes, eliminar el ruido y la redundancia.

Balancear el conjunto de datos:



Balancear el conjunto de datos:

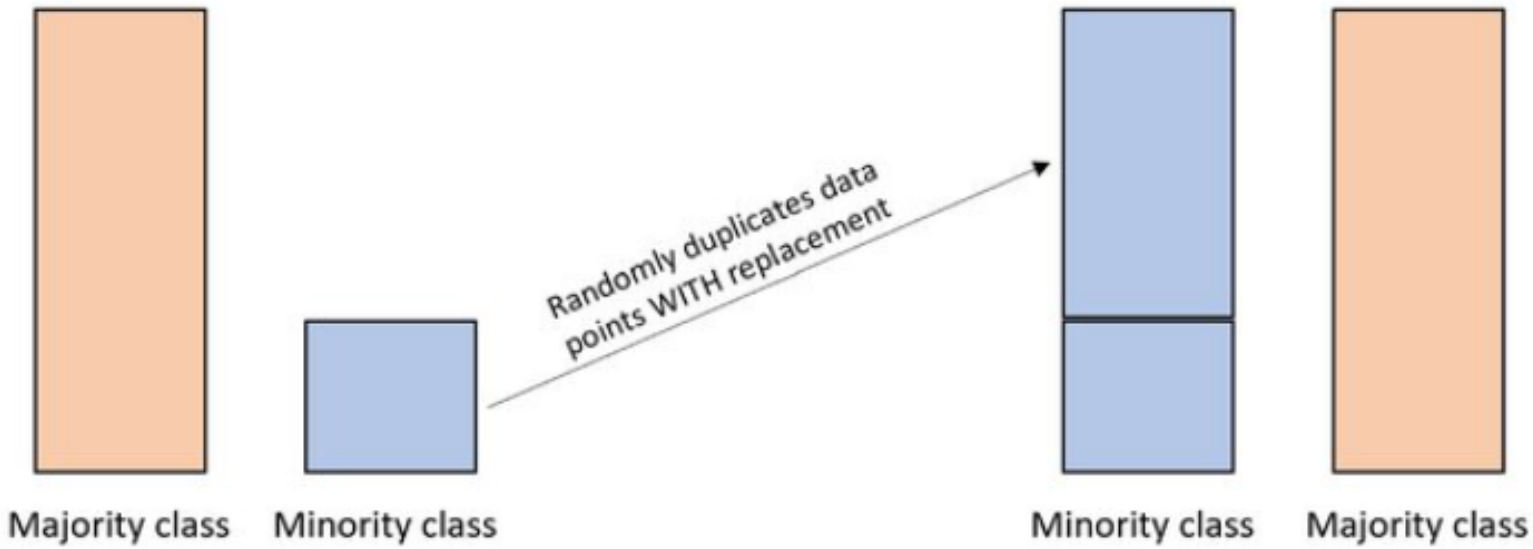


Tambien está la técnica de class weight, la cual asocia valores o pesos más altos a las muestras de la clase minoritaria y pesos más bajos a la clase mayoritaria durante el proceso de entrenamiento.

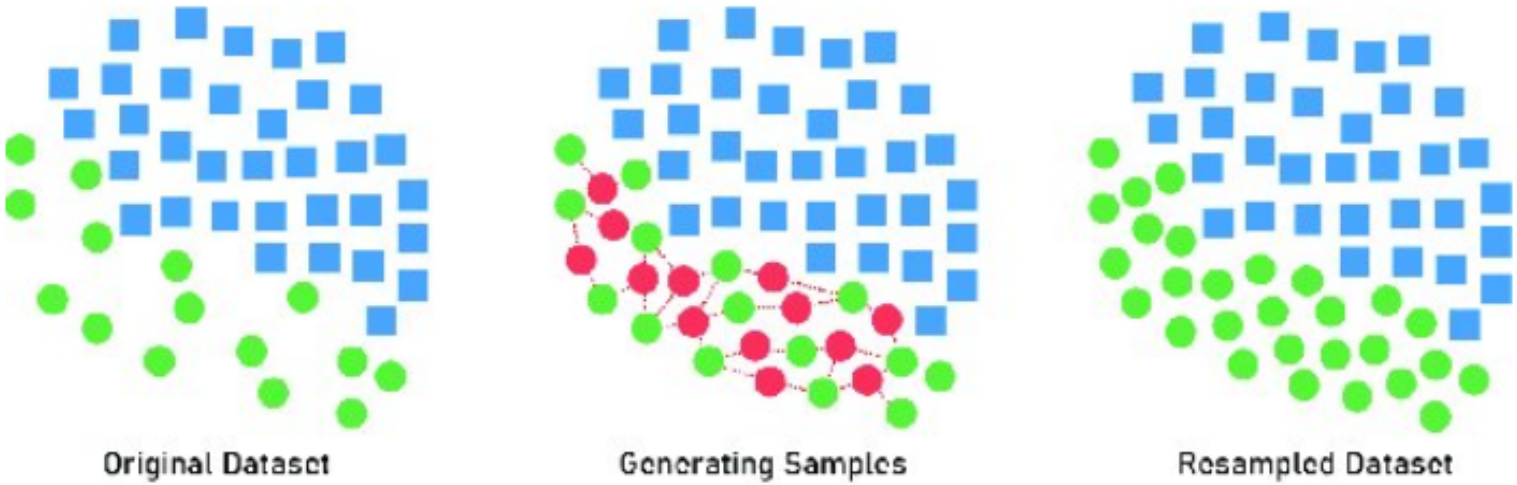
Balancear el conjunto de datos: Oversampling

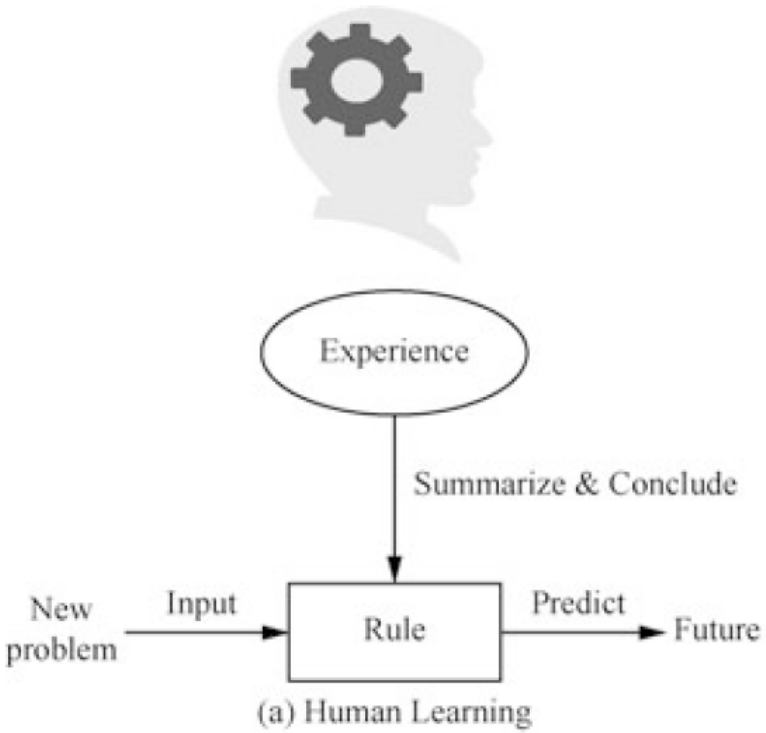
- Random Oversampling
- SMOTE

Random oversampling:

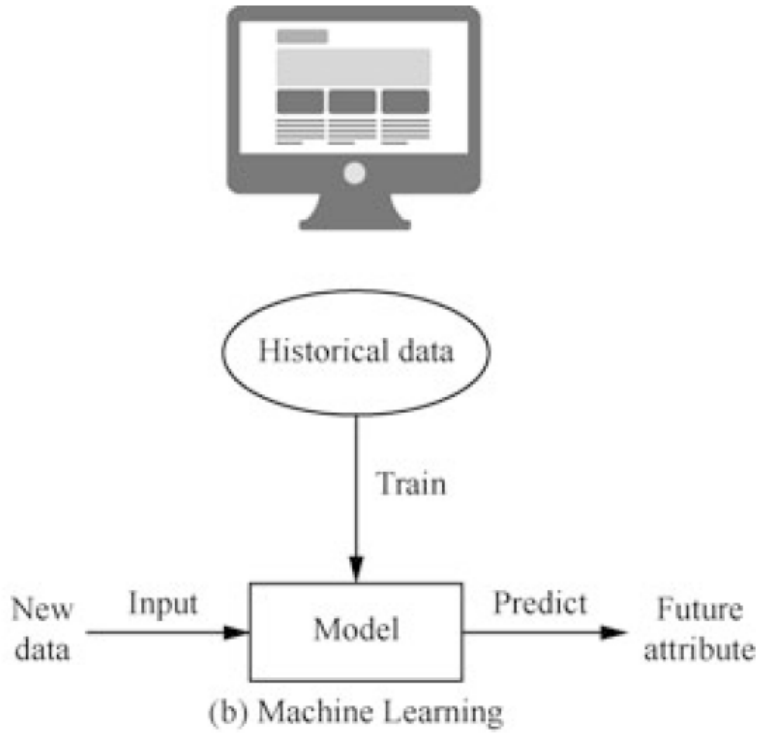


SMOTE: Synthetic Minority Oversampling Technique Las instancias nuevas no son meras copias de los casos minoritarios existentes. En su lugar, el algoritmo toma muestras del espacio de características de cada clase de destino y de sus vecinos más próximos. Luego, el algoritmo genera nuevos ejemplos que combinan las características del caso que nos ocupa con características de sus vecinos.





(a) Human Learning



(b) Machine Learning

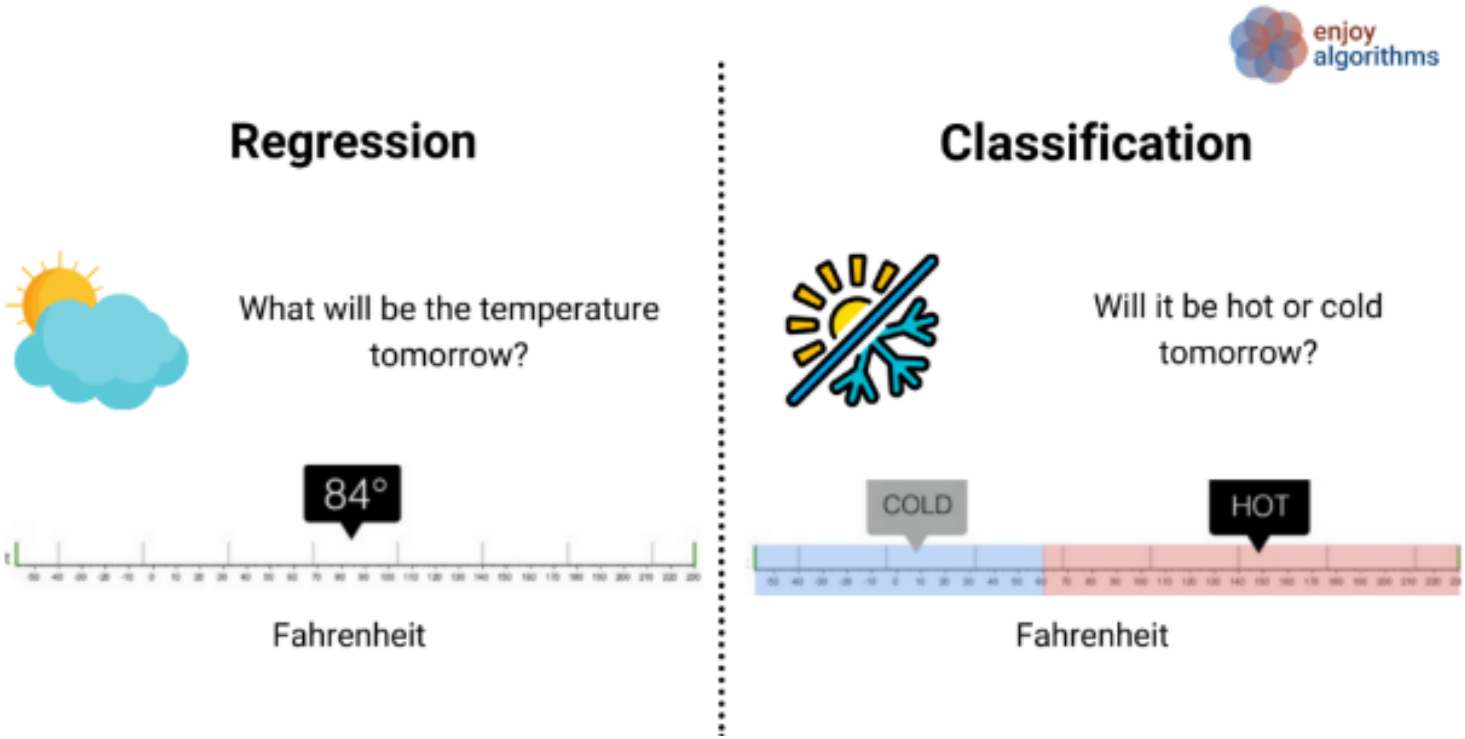
¿Cuáles son las 2 tareas principales que puede realizar los algoritmos de Machine Learning?

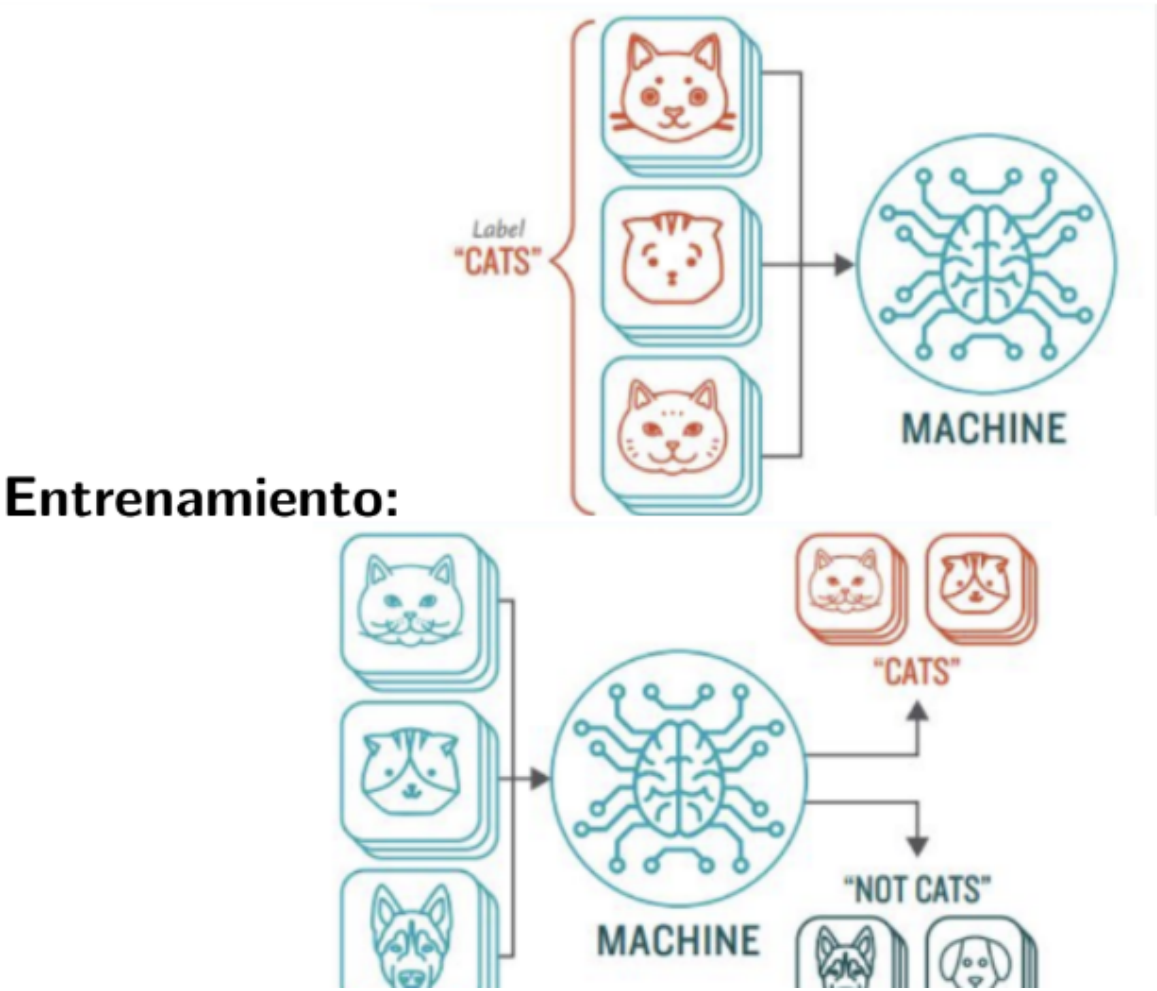
- **Clasificación:** Es el proceso de identificar un nuevo objeto/muestra de entrada no etiquetada como una clase.

- **Predicción:** Es el proceso de estimar un valor num erico o una salida continua en funcion del nuevo objeto/muestra de entrada no etiquetada.

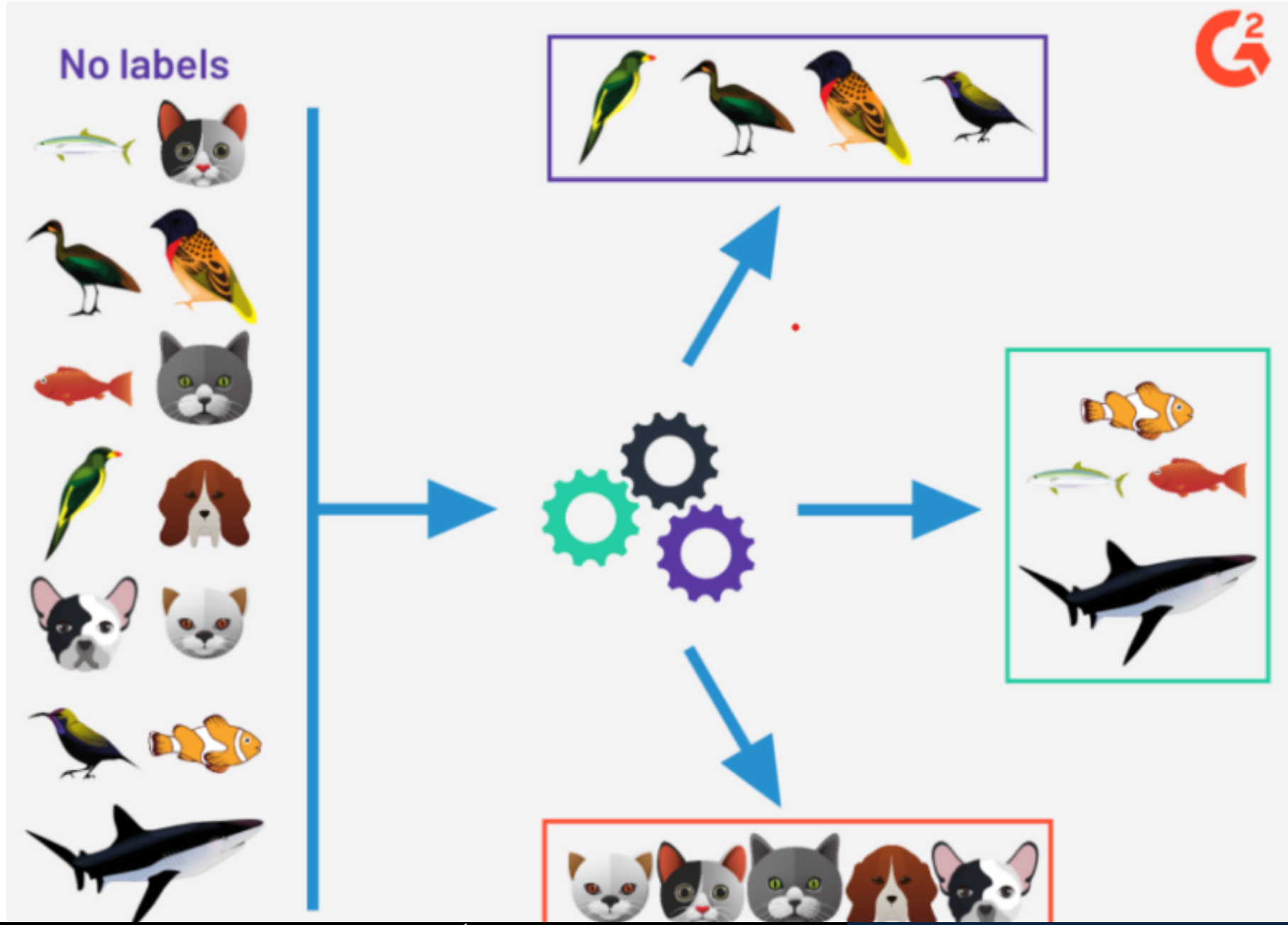
¿Cómo se le denominan a los algoritmos de ML que realizan tareas de clasificación? **Clasificadores**

¿Cómo se le denominan a los algoritmos de ML que realizan tareas de predicción? **Algortimos de regresión**


enjoy algorithms
Regression
What will be the temperature tomorrow?
84°
Fahrenheit
Classification
Will it be hot or cold tomorrow?
COLD
HOT
Fahrenheit


Label
"CATS"
MACHINE
Entrenamiento:
"CATS"
"NOT CATS"
MACHINE

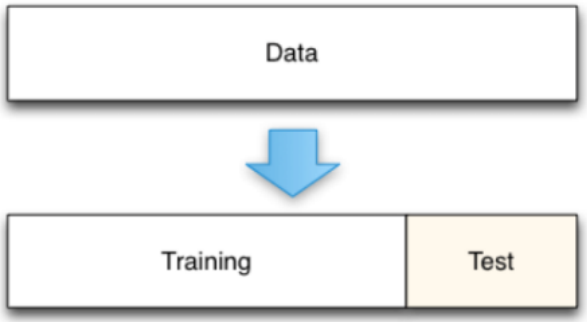
Técnicas de aprendizaje de ML:supervisado Y NO supervisad



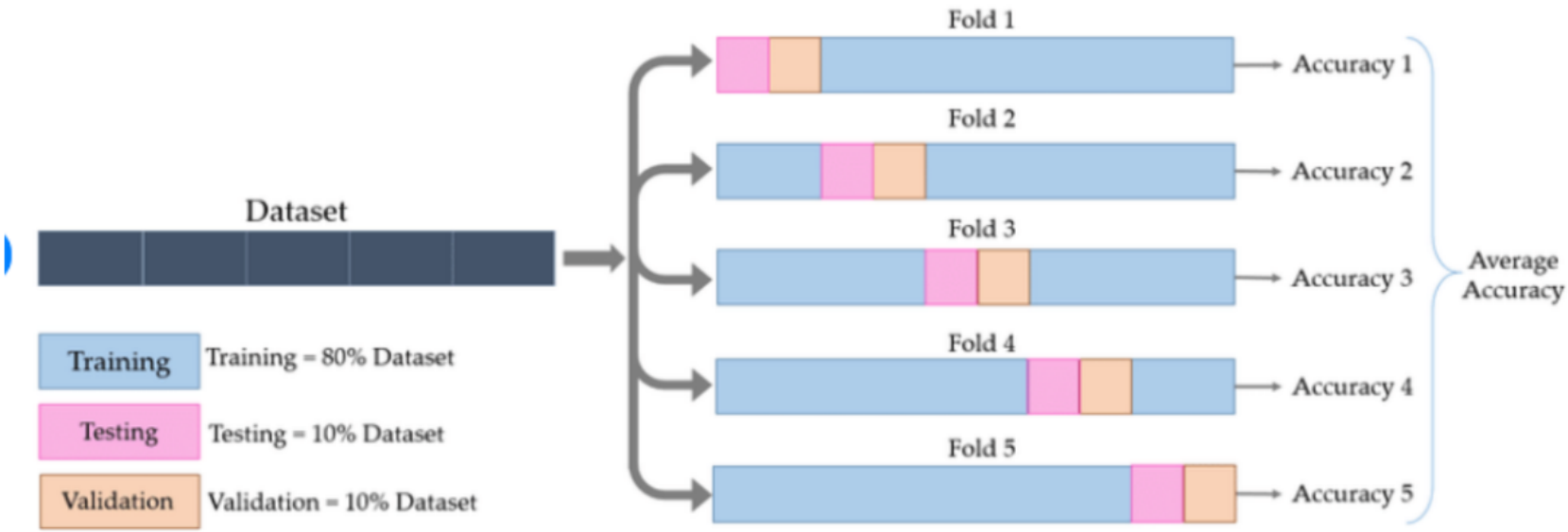
¿Para qué nos sirven los métodos de evaluación?

- La evaluación nos da evidencia para anticipar el correcto funcionamiento de los modelos de ML.
- Una evaluación sistemática es imprescindible para publicar resultados y avanzar en el estado del arte.

Dividir el dataset de forma aleatoria en dos subconjuntos o incluso tres subconjuntos.



K-fold cross validation.



Verdaderos positivos (TP-True positives): Mide la cantidad de datos identificados correctamente como la clase positiva.
Falsos positivos (FP-False positives): Mide la cantidad de datos identificados erróneamente como la clase positiva.
Falsos negativos (FN-False negatives): Mide la cantidad de datos identificados errónemante como la clase negativa.
Verdaderos negativos (TN-True negatives): Mide la cantidad de datos identificados correctamente como la clase negativa.

Matriz de confusión

Confusion Matrix		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

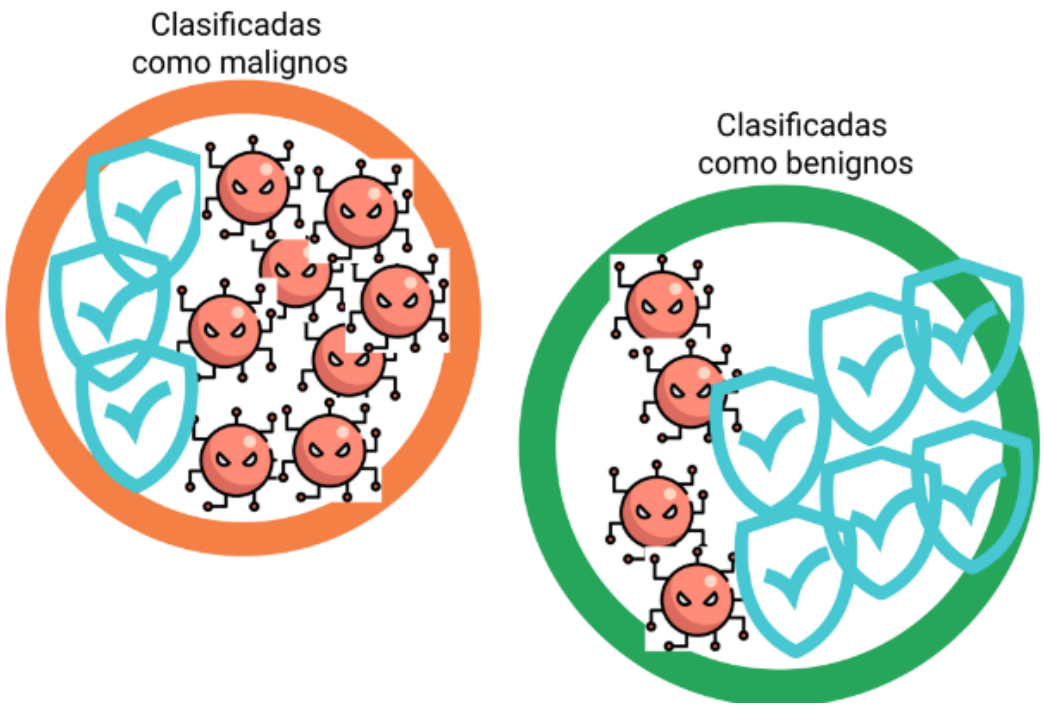
¿Qué pasa cuando tenemos más de dos clases?

Matriz de confusión

		PREDICTED classification			
	Classes	a	b	c	d
ACTUAL classification	a	TN	FP	TN	TN
	b	FN	TP	FN	FN
	c	TN	FP	TN	TN
	d	TN	FP	TN	TN

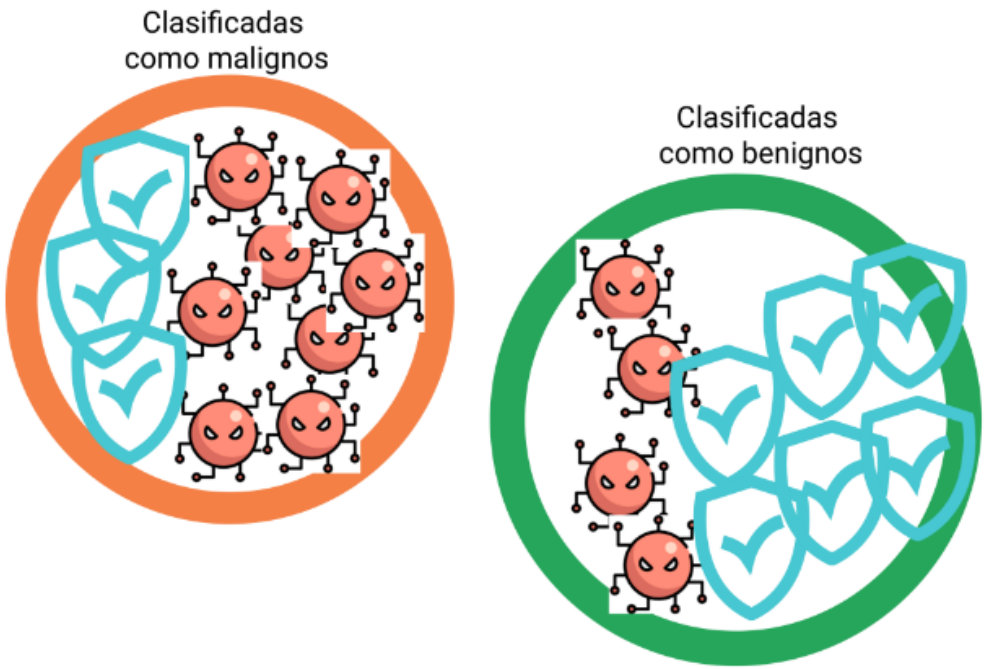
Accuracy/exactitud: describe la proporción de muestras identificadas correctamente de todas las instancias.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



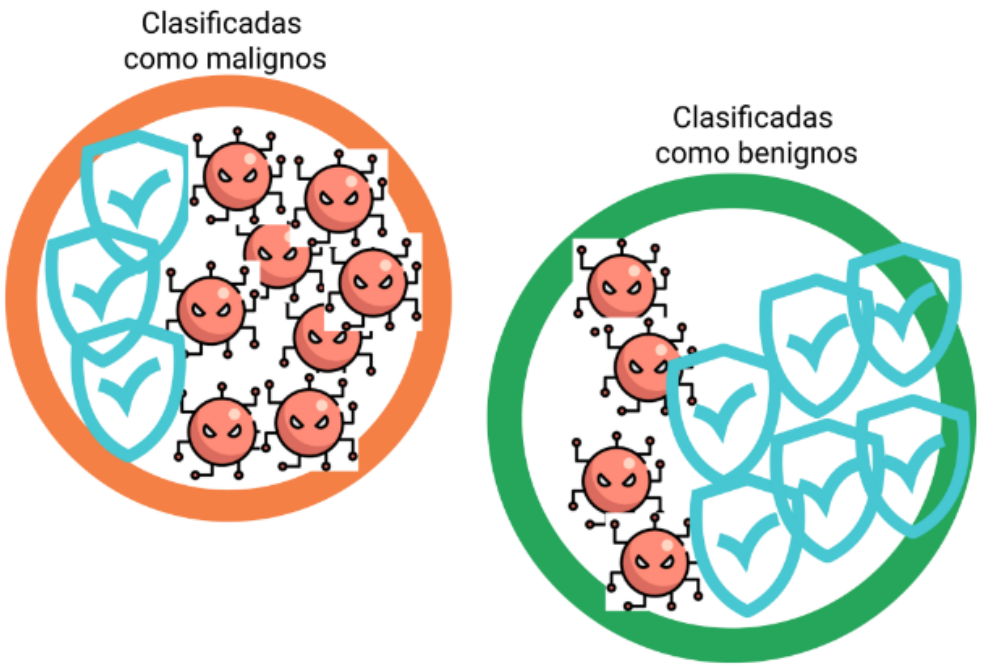
Precisión: Identifica la relación entre el número de muestras verdaderas y todas las observaciones pronosticadas como positivas.

$$Precisión = \frac{TP}{TP + FP} \tag{2}$$



Recall/Recuperación: calcula la relación entre el número total de verdaderos positivos y todos los verdaderos positivos.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$



F1-Score: Es un promedio armónico de métricas de Recall y precisión tomando su promedio ponderado.

$$F1 - score = 2x\frac{PrecisionxRecall}{Precision + Recall}$$

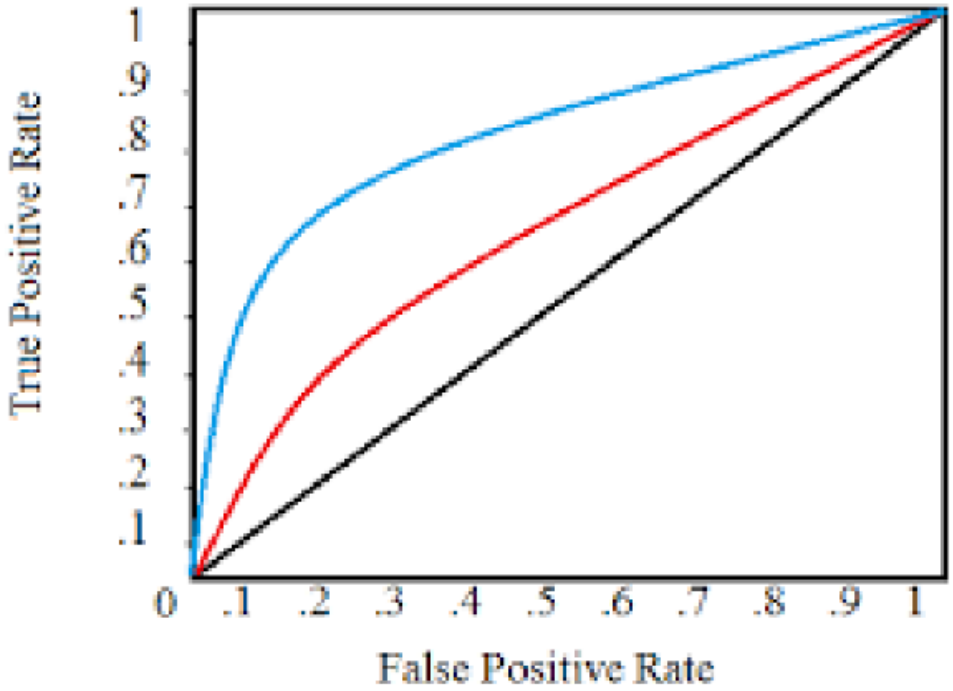
Razón de detección - Razon de verdaderos positivos (True Positive Rate- TPR): Corrresponde al número de muestras positivas reales identificadas correctamente en relación con el número total de muestras positivas.

$$TPR : \frac{TP}{TP + FN} \tag{5}$$

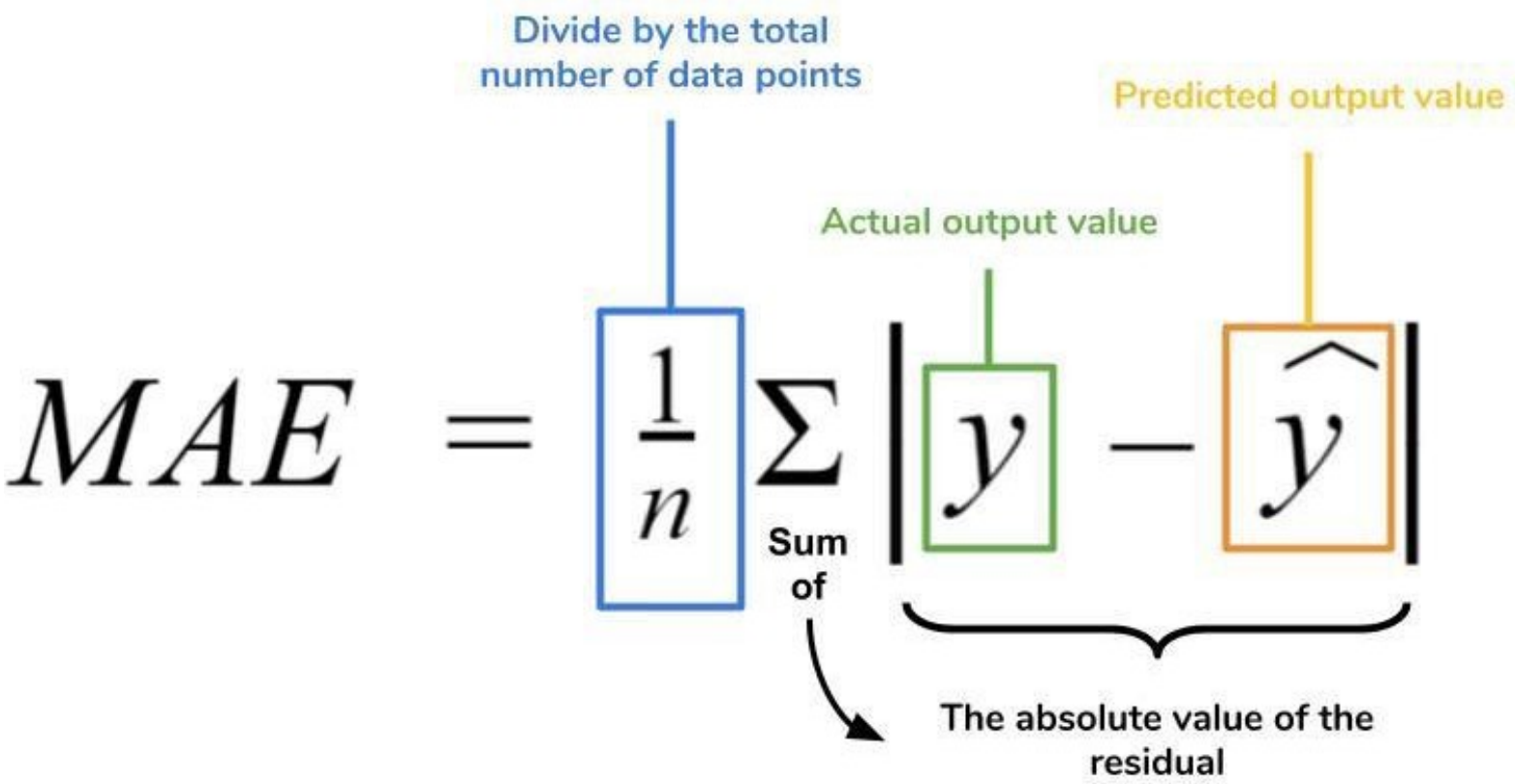
Razon de falsos positivos/falsas alarmas (False Positive Rate- FPR): Corrresponde al número de muestras negativas identificadas incorrectamente como positivas con relación con el total de instancias negativas.

$$FPR : \frac{FP}{FP + TN} \tag{6}$$

Curva ROC ((receiver operating characteristic curve)) y Área bajo la curva (AUC)



Error Absoluto Medio (MAE)



Error Cuadrático Medio (MSE)

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

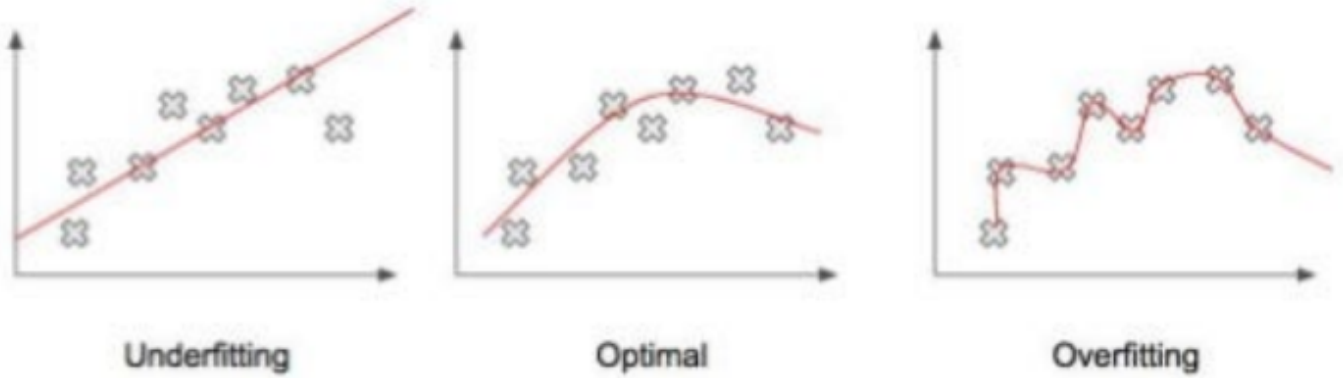
makes result quadratic

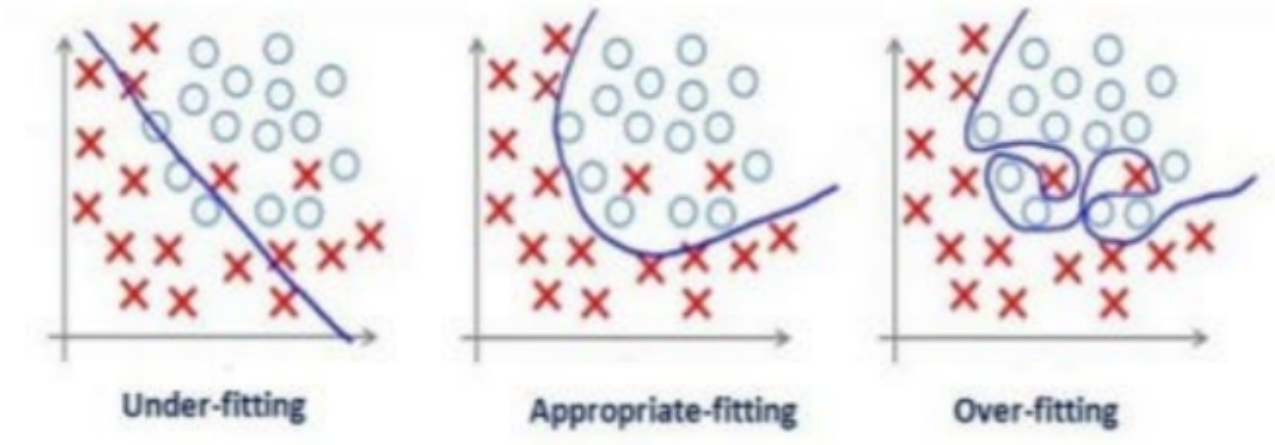
y_i: true y

$\hat{y}_i$: estimate of y

$\frac{1}{N}$: average over all results

Overfitting / Sobrajuste y Underfitting/Subajuste





- Descargar un dataset
- Familiarizarse con la libreria scikit-learn de python
- Identificar tarea a resolver
- Analizar el dataset de acuerdo a los puntos discutidos previamente.
- Elegir los pasos de pre-procesamiento más adecuados y justificarlos.
- Balancear el dataset (en caso de ser necesario)
- Guardar los cambios en cada modificación que se realice en el datset.
- Escribir un reporte en formato pdf redactando las actividades realizadas.
- Subir en classroom código y reporte.