

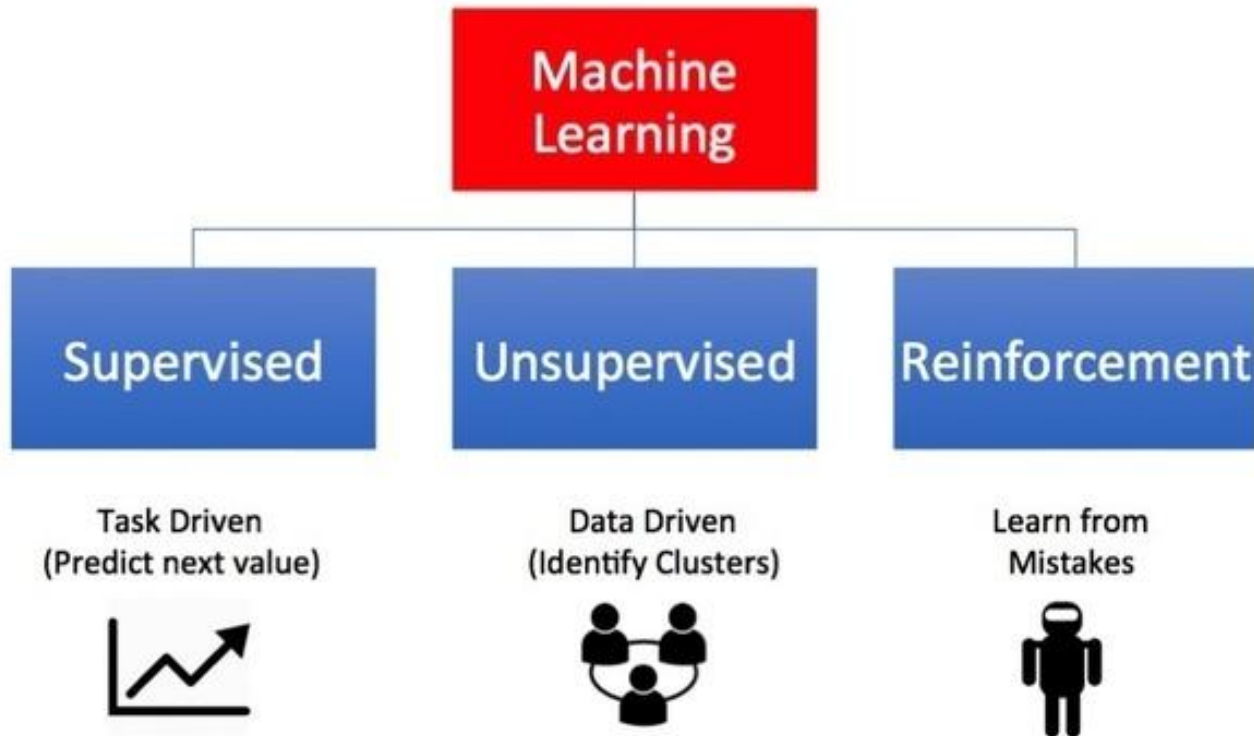




Fundamentos para el preprocesamiento de los datos

Sesión 2:

Types of Machine Learning

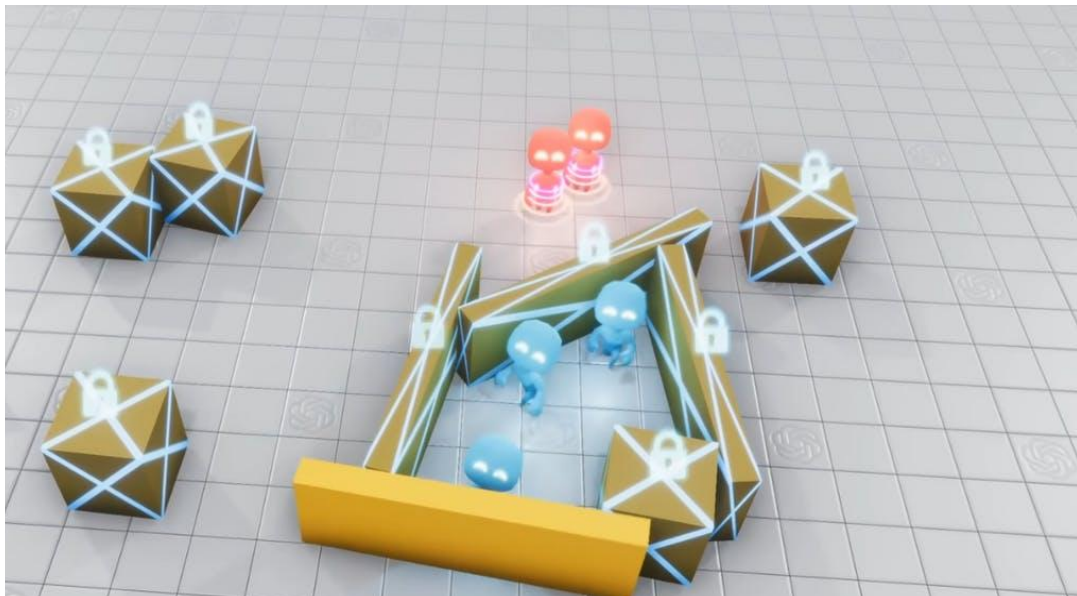


Aprendizaje Supervisado

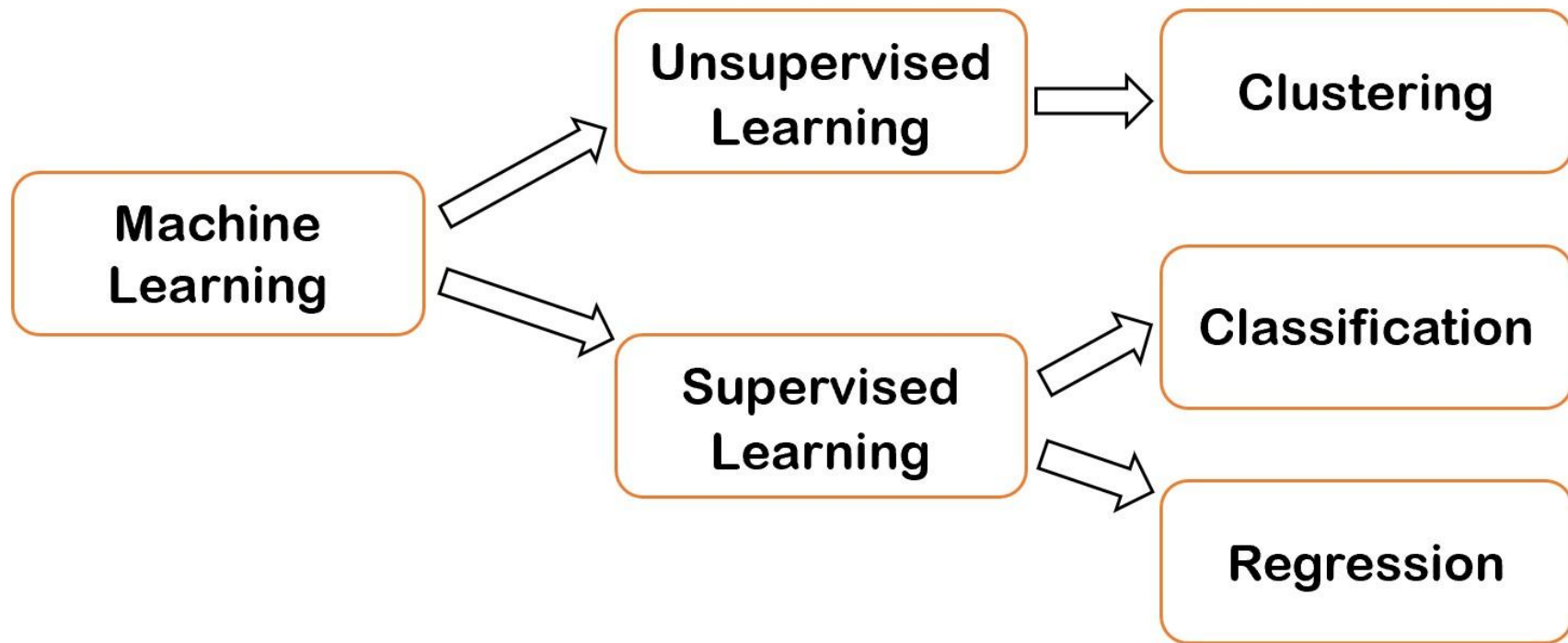


- Aprende de muestras que ya tienen respuestas dadas. Ya sea hecho por humanos o por fenómenos en la naturaleza.
- Tras mostrarle muchos ejemplos, el algoritmo debería ser capaz de predecir bien valores que no ha visto antes (osea no es simple memoria)
- Dale a estos algoritmos datos de entrada y de salida (target o variable objetivo) y si existe una relación es capaz de encontrarla.

Aprendizaje Reforzado



- Existen agentes diseñados para aprender de sus errores. Nadie les dice qué hacer, sino que los agentes mismos aprenden a “sobrevivir” a la función de coste.
- Este tipo de algoritmos son programados para hacer secuencias de acciones que pueden ser recompensadas o penalizadas.
- Dado el punto anterior, estos algoritmos son puestos en práctica en videojuegos. (como Dota2, Starcraft II, etc)



DATA CLEANING



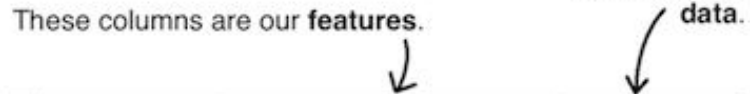
Pre-procesando los datos:

Conocer tus datos: Exploratory Data analysis

- Es necesario hacerse preguntas sobre qué es lo que significan y dicen nuestras variables.
- Pensar el problema en base a nuestros datos es crucial para este momento.
- Este conocimiento de los datos se puede llevar a cabo con un análisis estadístico exploratorio.

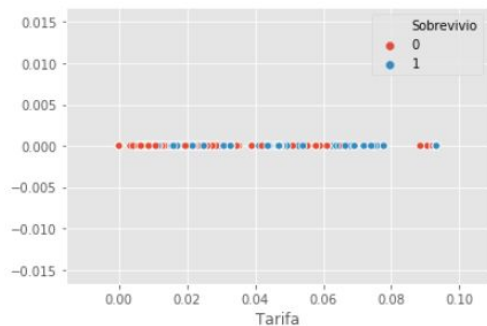
These columns are our **features**.

Because we know the survival status of each patient, this is **labeled data**.

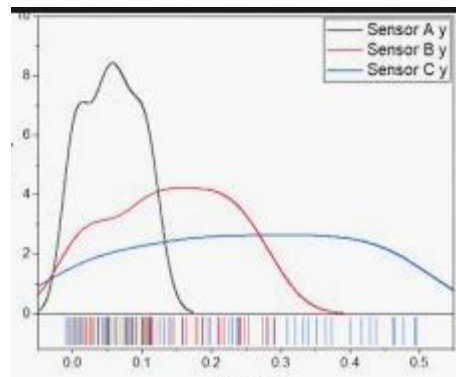
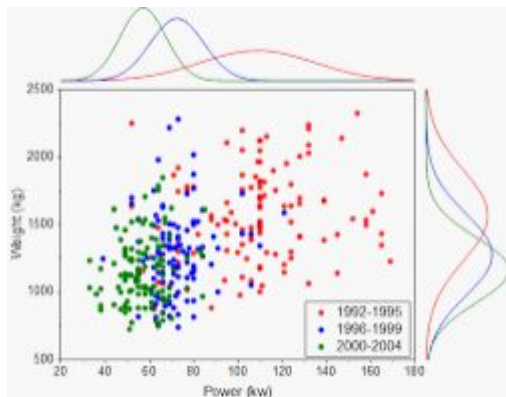


	patient_age	operation_year	nodes_detected	survival_status
1				
2	30	64	1	1
3	30	62	3	1
4	30	65	0	1
5	31	59	2	1
6	31	65	4	1
7	33	58	10	1
8	33	60	0	1
9	34	59	0	2
10	34	66	9	2
11	34	58	30	1
12	34	60	1	1
13	34	61	10	1
14	34	67	7	1
15	34	60	0	1
16	35	64	13	1
17	35	63	0	1
18	36	60	1	1
19	36	69	0	1
20	37	60	0	1

Análisis Univariado



Análisis Multivariado

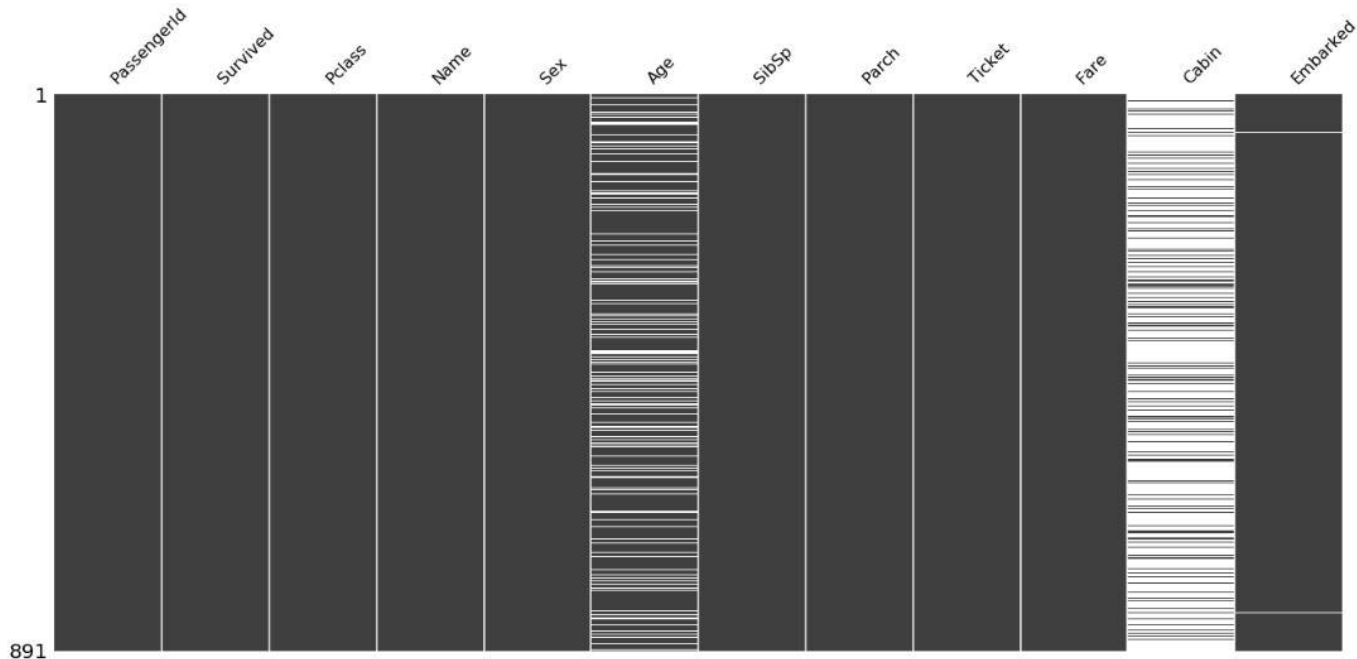




accelerate your learning

Encontrar y procesar valores nulos

(Find and process null values)

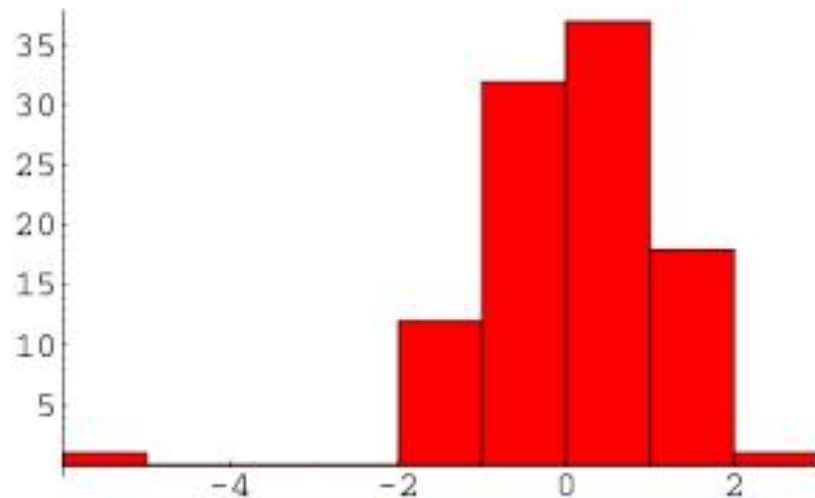
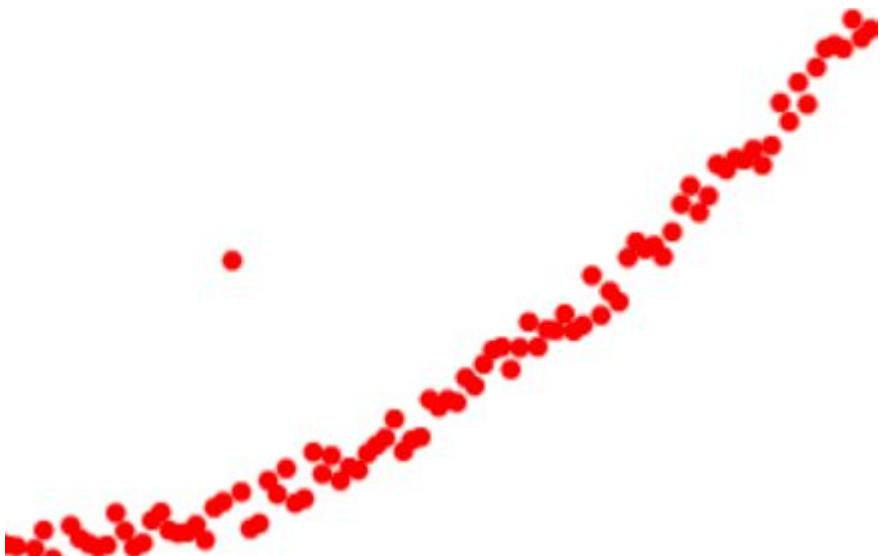


```
df.isnull().sum()
```

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2
dtype:	int64

Encontrar y procesar valores atípicos

(Find and process outlier values)





accelerate your learning

Ingeniería de características

(Feature Engineering)

Before				After Feature Engineering						
id	familyCnt	totalInc	totalExp	id	familyCnt	totalInc	totalExp	incPerPerson	expPerPerson	savingsPerPerson
101	2	68000	48000	101	2	68000	48000	34000	24000	10000
102	4	72000	66000	102	4	72000	66000	18000	16500	1500
103	3	34000	33000	103	3	34000	33000	11333.3	11000	333.3
104	3	44000	41000	104	3	44000	41000	14666.7	13666.7	1000
105	5	52000	50000	105	5	52000	50000	10400	10000	400

$\text{incPerPerson} = \text{totalInc} / \text{familyCnt}$; $\text{expPerPerson} = \text{totalExp} / \text{familyCnt}$; $\text{savingsPerPerson} = \text{incPerPerson} - \text{expPerPerson}$



accelerate your learning

Encodear variables categóricas

(Encoding categorical values)

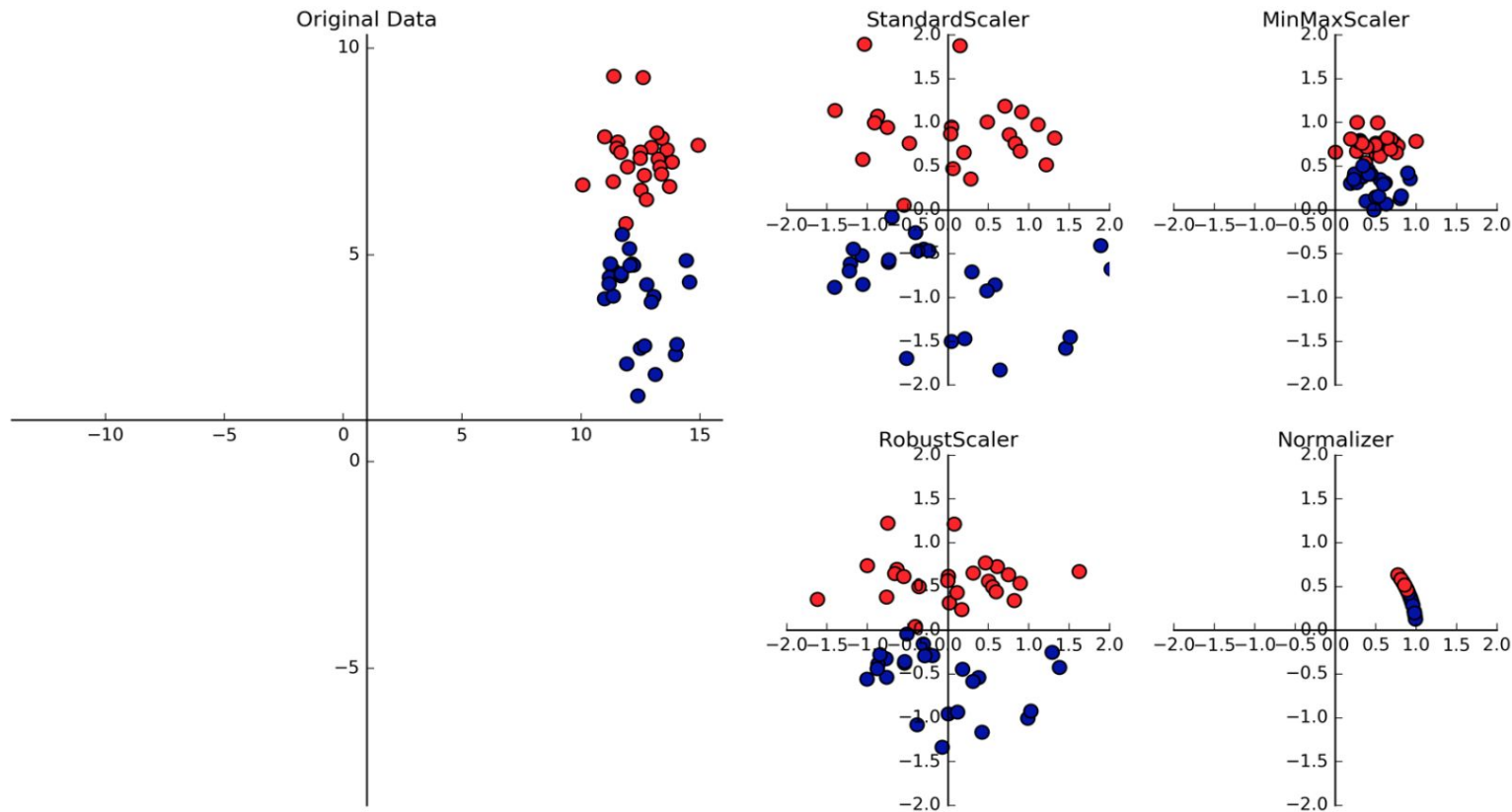
Before		After Label Encoding	
id	country	id	countryLabel
101	NZ	101	1
102	BR	102	0
103	US	103	2

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	male	22	1	0	7.2500	NaN	S
1	1	1	female	38	1	0	71.2833	C85	C
2	1	3	female	26	0	0	7.9250	NaN	S
3	1	1	female	35	1	0	53.1000	C123	S
4	0	3	male	35	0	0	8.0500	NaN	S

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	0	0.022728	1	0	0.004745	0
1	1	1	2	0.039257	1	0	0.046655	2
2	1	3	2	0.026860	0	0	0.005187	0
3	1	1	2	0.036158	1	0	0.034754	0
4	0	3	0	0.036158	0	0	0.005269	0

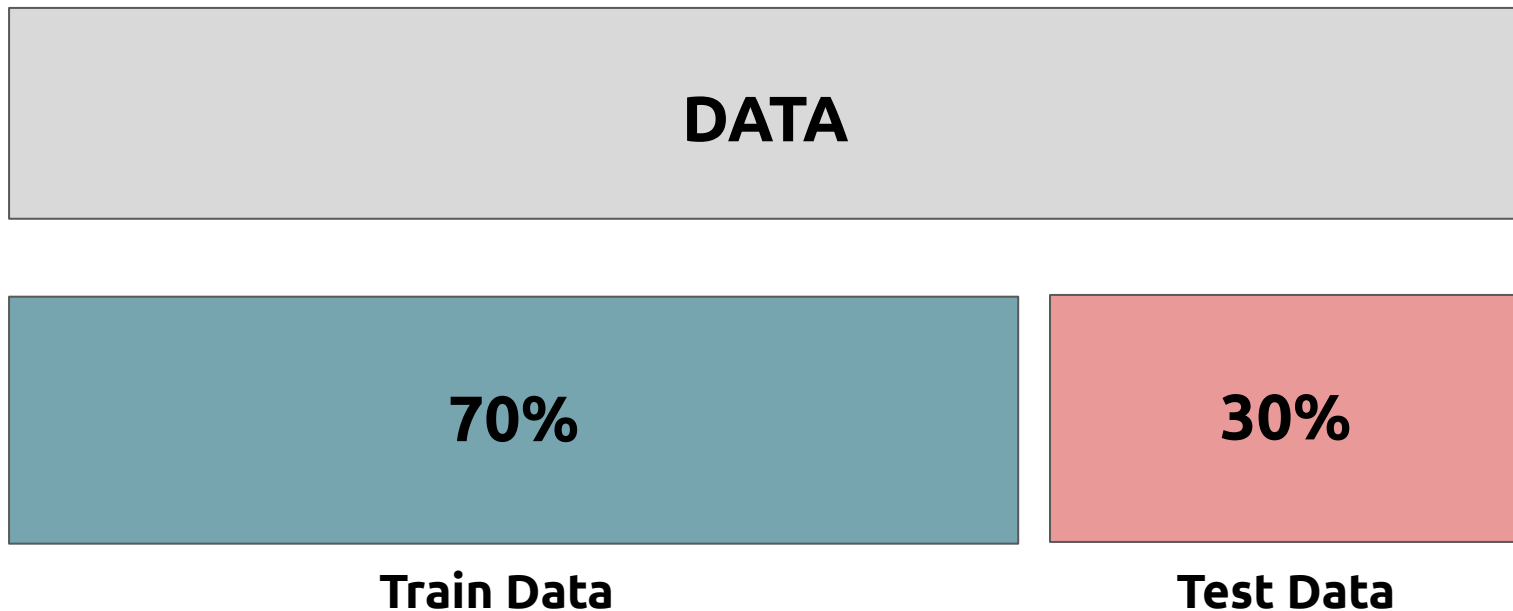
Escalamiento de los datos

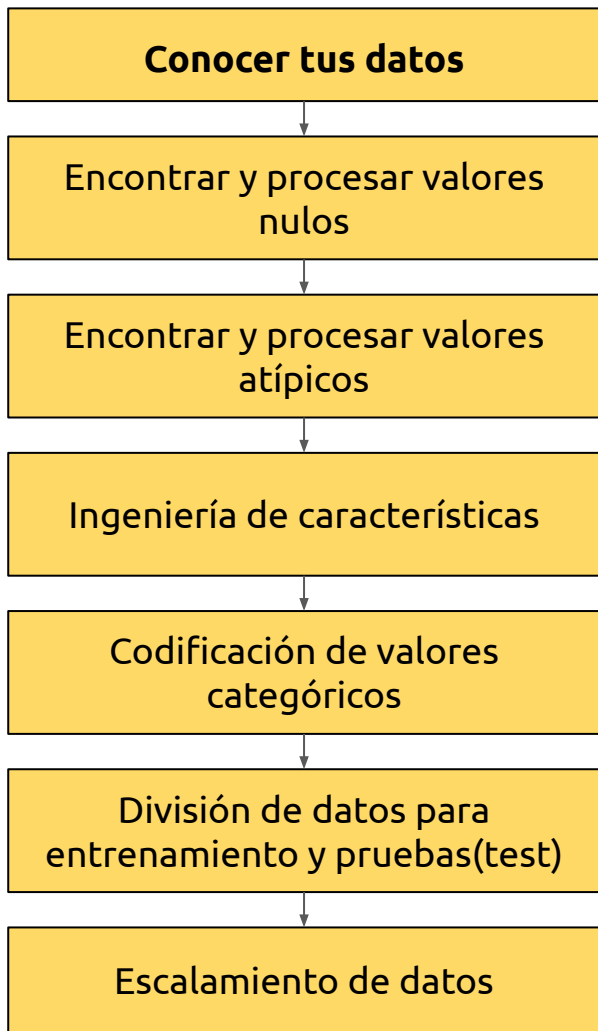
(Scaling data)



Datos de entrenamiento y de prueba

(Train data and test data)





- significado de cada atributo
- número de filas y columnas
- cantidad de atributos(columnas) con valores numéricos y categóricos
- histogramas de los valores de las columnas

RoadMap:

Preprocessing for Supervised Machine Learning



accelerate your learning

Normas de clase

- No se permiten bebidas (sin tapa) durante la clase para prevenir el daño a sus computadoras y el ambiente de estudios.
- Evitar consumir snacks que causen ruido durante clases o ensucien el ambiente
- Se brindará una tolerancia de 15 min para dar inicio a la clase, en caso de emergencia por favor comunicarse con los profesores con anticipación
- Se podrá admitir hasta un máximo de **dos (2) inasistencias** para obtener el certificado.

Normas de clase

- Todas las preguntas son bienvenidas, siempre que estén relacionadas al contenido de la clase en curso.
- No es bienvenida la burla o *bullying* entre alumnos y menos de parte de nuestros profesores.
- HackSpace es un ambiente de aprendizaje que busca fomentar el desarrollo **sin importar el sexo, raza, religión u orientación sexual**. Toda práctica de discriminación está prohibida dentro de este espacio.