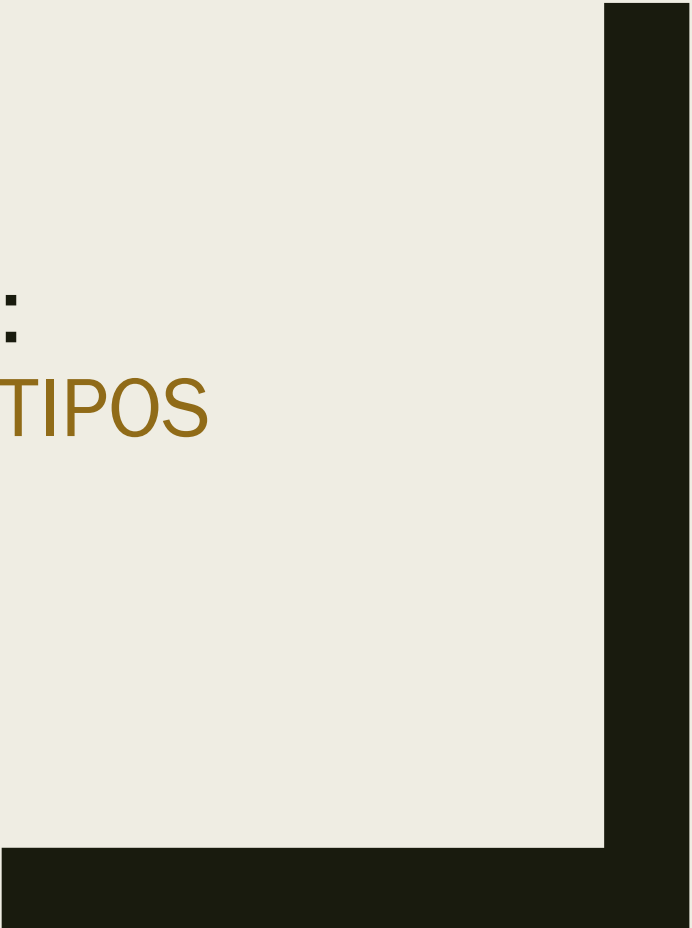


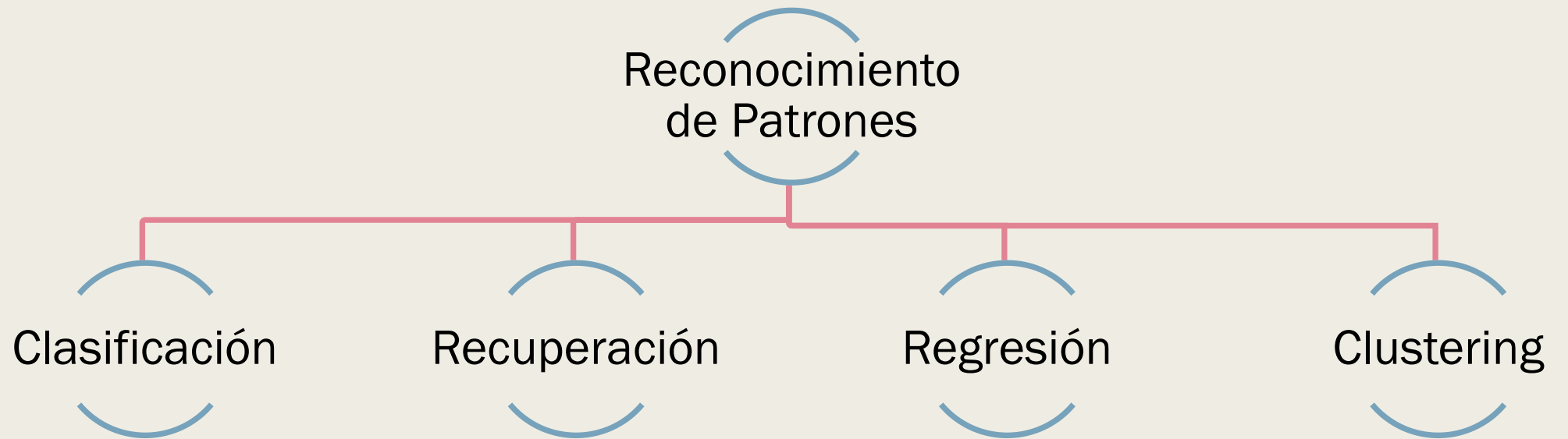


# ANALÍTICA AVANZADA DE DATOS: CONCEPTOS BÁSICOS

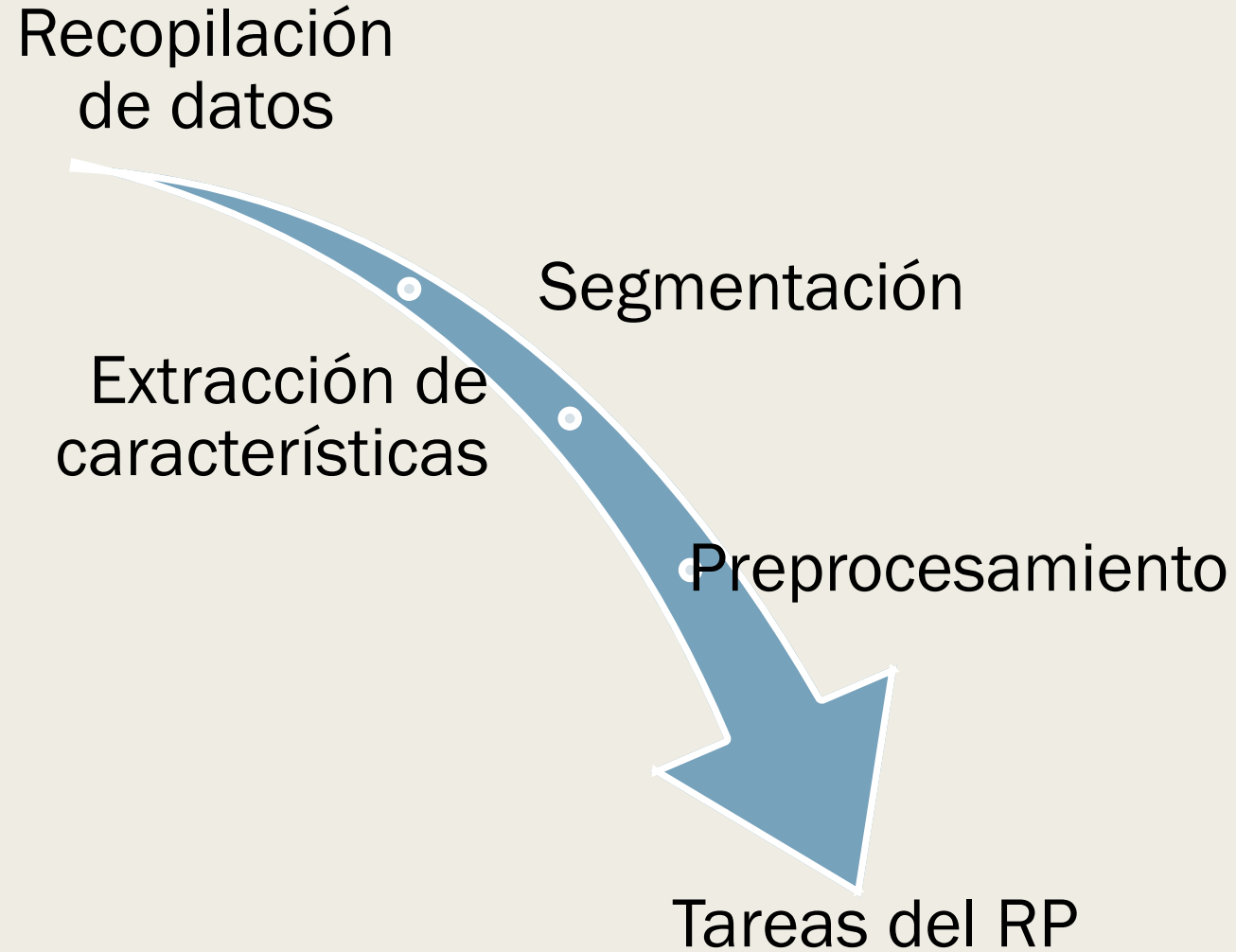
A. Alejandra Sánchez Manilla  
asanchezm.q@gmail.com



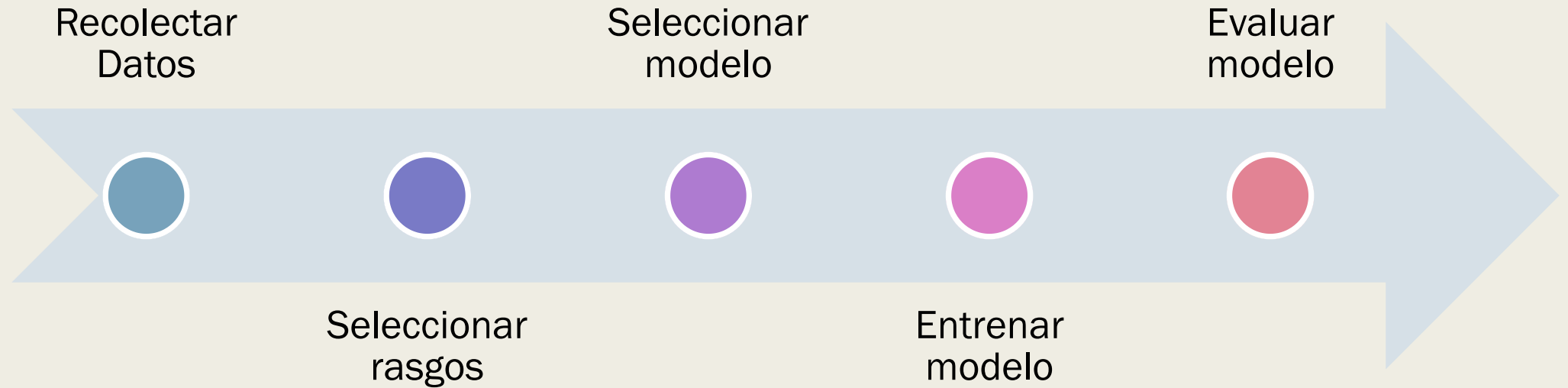
# Tareas del Reconocimiento de Patrones



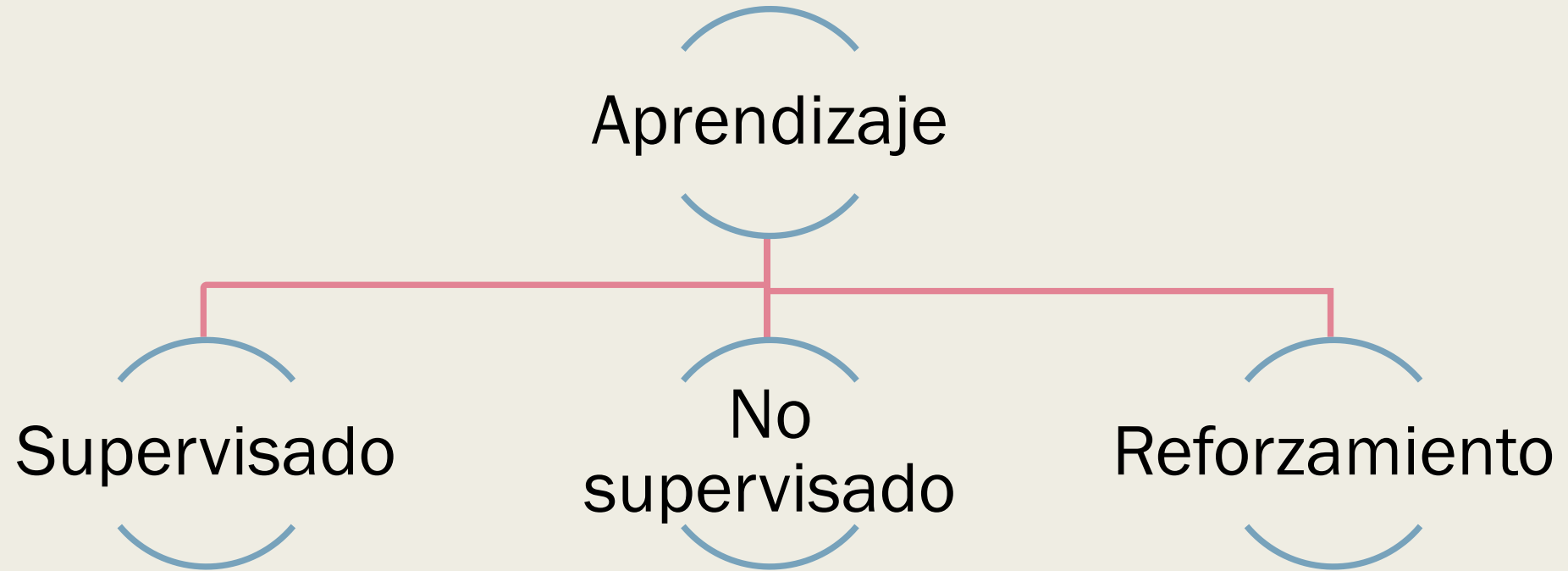
# Etapas en un sistema de RP



# Ciclo de Diseño



# Tipos de Aprendizaje



## Aprendizaje Supervisado

- Se proporciona un conjunto de datos, usado para entrenar al sistema RP, el cual consta de patrones acompañados de sus clases (valor objetivo).
- Es un enfoque dirigido por el concepto.
- Si el objetivo del sistema RP es asignar una categoría entonces estamos ante una tarea de clasificación. Si, por otro lado, la salida del sistema consiste un 1 o más valores continuos entonces estamos ante una tarea de regresión.

## Aprendizaje no Supervisado

- O clustering, no existe un experto. El sistema RP forma agrupaciones naturales basándose en los patrones de entrenada.
- Es un enfoque dirigido por los datos, cuyo objetivo es descubrir grupos de datos similares.
- Dado un determinado conjunto de datos o una función de costo, diferentes sistemas de *clustering* pueden conducirnos a diferentes agrupaciones de los datos.
- Generalmente el usuario debe suponer el número de cluster de antemano. ¿Cómo evitar una representación inadecuada de los datos?

## Aprendizaje por Reforzamiento

- El objetivo es encontrar acciones adecuadas en una situación dada de forma que se maximice una recompensa.
- En este enfoque no se da un valor objetivo, sino que se debe encontrar la solución mediante un proceso de prueba y error.
- Alcanzar un balance entre explotar las acciones conocidas que producen una recompensa y explorar nuevas acciones.
- **Ejemplo:** un robot debe decidir entre entrar a un cuarto a recolectar más basura o buscar una estación para recargar su batería.



## Notación Básica

Símbolo	Significado	Consideraciones
$A$	Conjunto genérico de donde se toman valores para las entradas de los patrones (vectores)	Ejemplos: $A = \{0,1\}$
$X$	Banco de datos (conjunto de patrones)	
$N$	Cardinalidad del conjunto de patrones	$N =  X $ , $N \in \mathbb{Z}^+$
$x^i$	$i$ -ésimo patrón del conjunto de patrones	$x^i \in X$ , $i = \{1,2,\dots,N\}$
$n$	Dimensión de los patrones, es decir número de rasgos o atributos	$n \in \mathbb{Z}^+$
$x_j^i$	La $j$ -ésima componente del patrón $i$	$i = \{1,2,\dots,N\}$ $j = \{1,2,\dots,n\}$
$c$	Número de clases	$c \in \mathbb{Z}^+$
$\omega_k$	$k$ -ésima clase	$k = \{1,2,\dots,c\}$

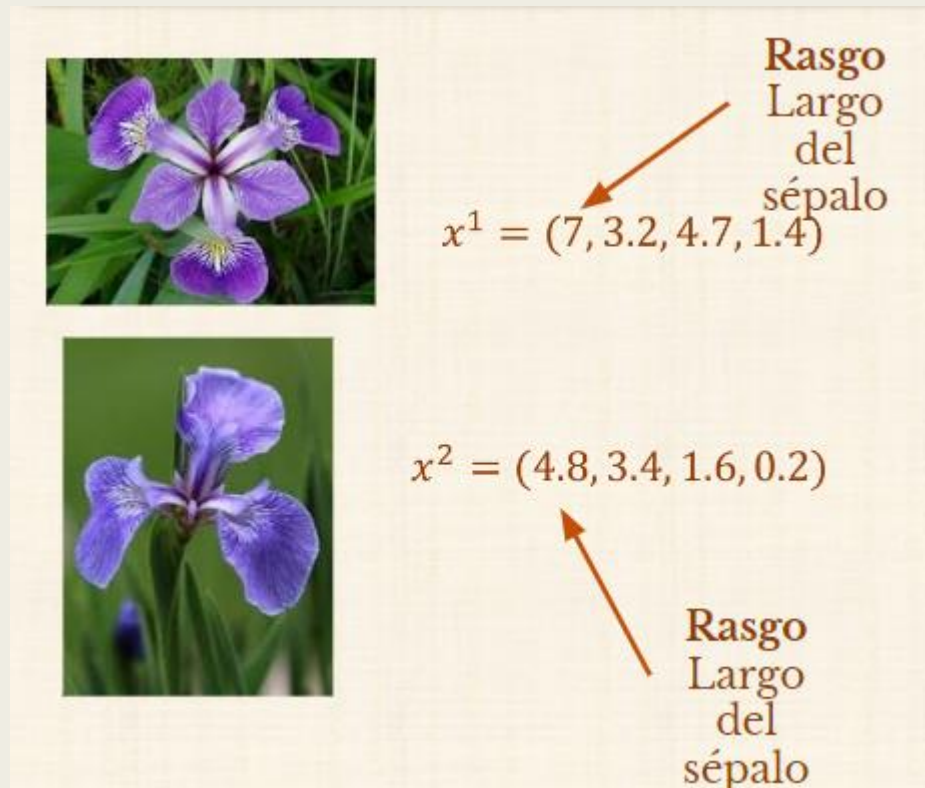
# Conceptos Básicos

## Notación Básica

Símbolo	Significado	Consideraciones
$E$	Conjunto de entrenamiento.	$E \cup P = X$ $E \cap P = \emptyset$
$P$	Conjunto de prueba.	
$N_E$	Cardinalidad del conjunto de entrenamiento	
$N_P$	Cardinalidad del conjunto de prueba	
$N_E(\omega_k)$	Número de patrones de la clase k en el conjunto de entrenamiento	
$N_P(\omega_k)$	Número de patrones de la clase k en el conjunto de prueba	
$\tilde{x}$	Patrón cuya clase se desconoce.	No es necesario agregar un índice, salvo que se requiera un ordenamiento.

# Conceptos Básicos

**Patrón:** un conjunto de rasgos que representan un objeto o problema. También se le conoce como *instancia* o *vector característico*



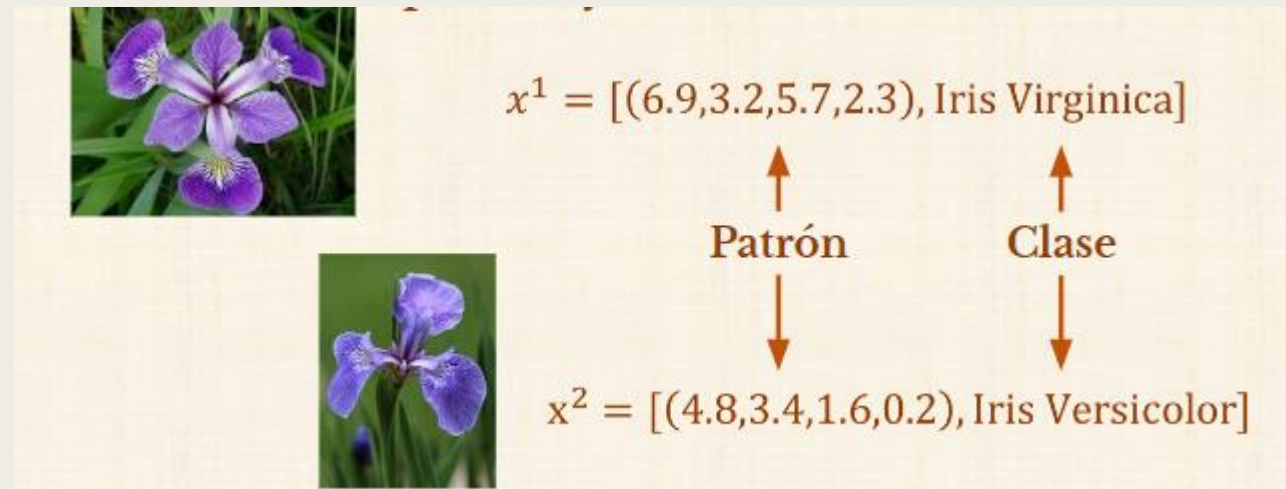
Nos referimos a cada rasgo del patrón con la siguiente notación  $x_j^i$  lo que significa el  $j$ -ésimo rasgo del  $i$ -ésimo patrón

# Conceptos Básicos

## Clases:

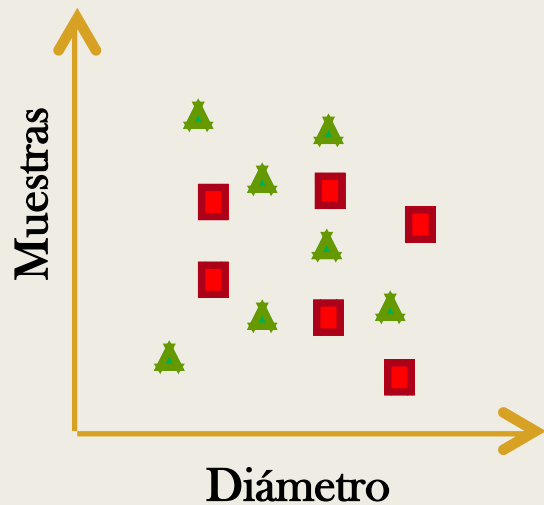
- *Es la categoría a la que pertenece un patrón*
- *La etiqueta que se le asigna a un patrón*

Se considera que los ejemplos que se le presentarán a un algoritmo de aprendizaje, para su entrenamiento, son un par [entrada-salida], donde la entrada es el patrón y la salida es la clase.

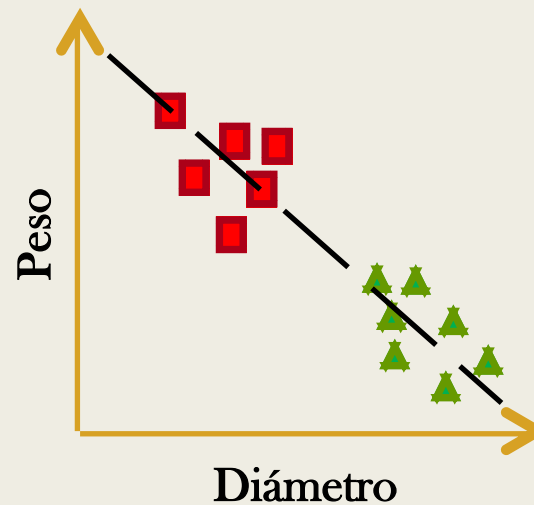


## Importancia de los atributos

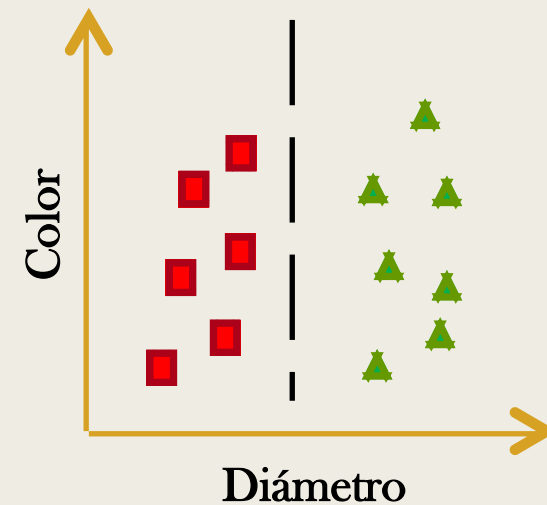
¿Qué atributos son más útiles? (Criterio de eficacia)



No es suficiente



Correlacionadas



Color: NO  
Diámetro: SI

## Importancia de los atributos

Los rasgos adecuados nos permiten diferenciar un patrón (objeto) de otro.

Una buena selección de atributos nos permite:

- *Enfocarnos en información relevante.*
- *Reducción de datos*
- *Mejoras en los resultados.*

**Problema:** generalmente no sabemos cual seleccionar, que representan y como ajustarlos.

## Banco de datos (dataset)

### Iris Data Set

*Download:* [Data Folder](#), [Data Set Description](#)

**Abstract:** Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	449872

# Conceptos Básicos

## Banco de datos

Un banco de datos esta generalmente organizado como una matriz de  $N$  filas (patrones)  $\times n$  columnas (rasgos), con una columna extra para la clase.

	Largo del Sépalo ( $n_1$ )	Ancho del Sépalo ( $n_2$ )	Largo del Pétalo ( $n_3$ )	Ancho del Pétalo ( $n_4$ )	Clase ( $n_5$ )
$N_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$N_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$N_3$	4.6	3.2	1.4	0.2	Iris-setosa
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N_{150}$	7.7	3.8	6.7	2.2	Iris-virginica



## Tipos de atributos (numéricos)

- Un atributo numérico es aquel que toma valores que se encuentran en el dominio de los números enteros o reales.
- Ejemplos:
  - edad ( $\mathbb{N}$ )
  - largo del pétalo ( $\mathbb{R}^+$ )
  - temperatura ( $\mathbb{R}$ )

*Además:*

- **Discretos:** toman valores finitos o contables.
- **Continuos:** aquellos que toman valores reales.
- **Binarios:** tipo especial de atributo discreto que solamente toma valores de 0 y 1.

## Tipos de atributos (categóricos)

Un atributo categórico es aquel cuyos valores son tomados de un conjunto de símbolos o cadenas.

Ejemplos:

- sexo (F,M)
- estado civil (Soltero, Casado, Viudo)

Además:

- **Nominal:** son datos que no tienen un orden, por tanto, solo operaciones de comparación (igualdad) tienen sentido. Ejemplo: estado civil.
- **Ordinal:** son datos que si representan un orden, por tanto, aquí si es lógico el uso de comparaciones de igualdad o desigualdad (mayor o menor). Ejemplo: nivel de educación (primaria, secundaria, posgrado).

## Tipos de atributos (representación)

- Si se considera un banco de datos con  $n$  atributos numéricos cada punto puede representarse como una tupla:

$$x^i = (x_1^i, x_2^i, \dots, x_n^i) \in \mathbb{R}^n$$

- O como un vector columna:

$$x^i = \begin{pmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_n^i \end{pmatrix} = (x_1^i, x_2^i, \dots, x_n^i)^T \in \mathbb{R}^n$$

## Clases

- Desde el punto de vista del problema de clasificación, los problemas pueden dividirse en:
  - **Biclases:** solo existen 2 categorías para los patrones en el dominio del problema. Ejemplo: problemas médicos, pacientes sanos o enfermos
  - **Multiclase:** existen más de 2 categorías para los patrones en el dominio del problema. Ejemplo: detección de intrusiones en una red, diferentes tipos de ataque

# Conceptos Básicos

## Clases – Problema biclase

- **Clases exclusivas:** son solo 2 clases y un objeto solo puede tener asignada una clase/categoría

Clase: imagen color



Clase: imagen blanco/negro



No puede ser ambas al mismo tiempo.

# Conceptos Básicos

## Clases – Problema multiclase

- **Clases no exclusivas:** un objeto puede tener múltiples clases/categorías asignadas (multietiqueta)



The diagram illustrates non-exclusive classes using two examples. The first example shows a landscape photo of a lake and mountains, which is labeled with multiple tags: Paisaje, Lago, Montañas, and Arboles. The second example shows a Netflix movie page for 'Despicable Me', which is labeled with multiple genres: Familiar, Niños, Comedia, and Animada. Arrows point from the images to their respective lists of labels.

**Etiquetas**

- Paisaje
- Lago
- Montañas
- Arboles

**Género**

- Familiar
- Niños
- Comedia
- Animada

## Clases – Problema multiclase

- **Clases exclusivas:** un objeto solo puede tener asignada una clase/categoría



# Conceptos Básicos

## Complejidad de Datos - Valores perdidos

- Este problema se presenta cuando el valor para al menos un rasgo de un patrón en el conjunto de datos no se encuentra presente.

Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			-
Carol		3	127	11	1	veggie	1	****
Mandy	44	2	68	8	1	chicken		null

- Causas:** mal funcionamiento de equipos de medición, cambio en el diseño durante la captura de datos, imposibilidad de coleccionar los datos, entre otras.



# Conceptos Básicos

## Valores perdidos ¿Cómo lidiar con ellos?

- Eliminar los patrones cuyos atributos contienen valores perdidos.

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.5	1.4	0.2	Setosa
4.6	3.1	NaN	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7		4.1	1.3	Versicolor
6.2	3.4	5.4	2.3	Virginica

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.5	1.4	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
6.2	3.4	5.4	2.3	Virginica

# Complejidad de Datos - ¿Cómo lidiar con ellos?

## Imputación

- Si el atributo que tiene valores perdidos es **numérico** se *rellenan los valores con el valor promedio o la mediana del atributo*
- Si el atributo que tiene valores perdidos es **categorico** se *rellenan los valores con la moda del atributo*

	col1	col2	col3	col4	col5		col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	mean() →	0	2.0	5.0	3.0	6.0	7.0
1	9	NaN	9.0	0	7.0		1	9.0	11.0	9.0	0.0	7.0
2	19	17.0	NaN	9	NaN		2	19.0	17.0	6.0	9.0	7.0

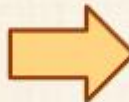
`sklearn.impute.SimpleImputer` donde se puede usar mean, median, constant y most\_frequent

# Complejidad de Datos - ¿Cómo lidiar con ellos?

## Imputación

- Se pueden diseñar estrategias más elaboradas para la imputación.

Ej.: Calcular media por categorías de algún rasgo



Estado	Salario	Años de experiencia
NY	57,400	10
TX	45,000	7
NJ	52,300	9
TX	39,500	5
NY	35,800	4
TX		6
NY	55,600	9

Estado	Salario	Años de experiencia
NY	57,400	10
TX	45,000	7
NJ	52,300	9
TX	39,500	5
NY	35,800	4
TX	42,250	6
NY	55,600	9

El promedio de todos los registros de TX

# Complejidad de Datos - ¿Cómo lidiar con ellos?

## Imputación

- Se pueden diseñar estrategias más elaboradas para la imputación.  
Ej.: Calcular media por categorías de la clase

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.5	1.4	0.2	Setosa
4.6	3.1		0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.8	4.1	1.3	Setosa
6.2	3.4	5.4	2.3	Versicolor

# Complejidad de Datos - ¿Cómo lidiar con ellos?

## Imputación

- Se pueden diseñar estrategias más elaboradas para la imputación.  
Ej.: Calcular media por categorías de la clase

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.5	1.4	0.2	Setosa
4.6	3.1	2.7	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.8	4.1	1.3	Setosa
6.2	3.4	5.4	2.3	Versicolor

El promedio de todos  
los registros de Setosa

# Complejidad de Datos - ¿Cómo lidiar con ellos?

## Uso de algoritmos:

- Usar algoritmos que sean robustos en el manejo de valores perdidos: KNN, Random Forest o GradientBoostingClassifier
- Usar un algoritmo de clasificación o regresión para realizar la imputación

## Complejidad de Datos - Valores atípicos (outliers)

- Los patrones atípicos son datos que presentan una diferencia significativa del resto de elementos en un conjunto de datos o en una clase en particular.
- Causas:
  - Malas mediciones al capturar los datos.
  - Mal etiquetado del patrón al asignarle una clase.
  - Características propias del concepto que se aprende.

## Patrones atípicos ¿Cómo lidiar con ellos?

1. Eliminar patrones atípicos que presenten valores evidentemente imposibles, un ejemplo sería edades negativas.
2. Tratar de *normalizar* los datos.
3. Calcular la *desviación estándar* y seleccionar datos que se alejen un determinado número de desviaciones estándar de la media del atributo que se analiza para ser eliminados del conjunto de datos.
4. Usar algoritmos que se vean menos afectados por valores atípicos, por ejemplo los Random Forest.
5. Métodos de proximidad como clustering, densidad o vecinos mas cercanos.



# Conceptos Básicos

## Patrones atípicos ¿Cómo lidiar con ellos?

1. Eliminar registros con valores evidentemente imposibles:

Edad	Sexo	Glucosa	Colesterol	Clase
56	F	1.4	2.3	Sano
46	M	5.7	5.7	Enfermo
75	F	4.9	8.9	Enfermo
-41	M	2.1	1.3	Sano
25	M	1.3	2.3	Sano

Edad	Sexo	Glucosa	Colesterol	Clase
56	F	1.4	2.3	Sano
46	M	5.7	5.7	Enfermo
75	F	4.9	8.9	Enfermo
25	M	1.3	2.3	Sano

# Conceptos Básicos

## Patrones atípicos ¿Cómo lidiar con ellos?

Normalizar y Estandarización:

- *El escalado de rasgos es uno de los pasos de procesamiento de datos más importantes en el aprendizaje automático*
- *Los algoritmos calculan la distancia entre los rasgos que están sesgados hacia valores numéricamente más grande si los datos no están escalados*

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.5	1.4	0.2	Setosa
40	3.1	1.2	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.5	4.8	1.3	Versicolor
6.2	3.4	5.4	50	Virginica

## Patrones atípicos ¿Cómo lidiar con ellos?

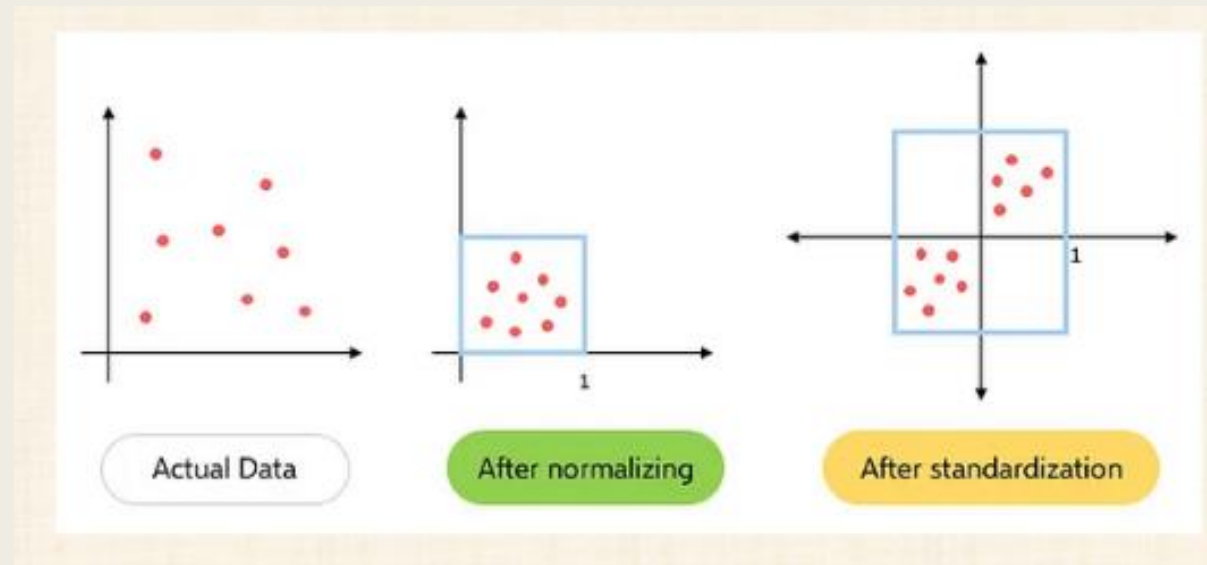
Normalizar y Estandarización:

Normalización	Estandarización
Valores mínimos o máximos son usados para el escalado de datos.	La media y la desviación estándar son usados para el escalado de datos.
Escala los valores en el rango $[0,1]$ o $[-1,1]$	No está limitado a un rango.
Se ve afectado por valores atípicos.	Se ve menos afectado por valores atípicos.
Es útil si no sabemos la distribución de los datos.	Es útil cuando la distribución de los datos es normal o gaussiana.
Llamada normalización por escala.	Llamada normalización Z-score

## Patrones atípicos ¿Cómo lidiar con ellos?

Normalizar y Estandarización:

- **Normalización:** escalar los datos
- **Estandarización:** escalar la distribución (distribución normal)



## Desbalance de Clases

Este fenómeno se presenta cuando el número total de instancias en una clase (clase mayoritaria) es significativamente mayor que el número de instancias de otra clase (clase minoritaria).

$$IR = \frac{\# \text{ instancias clase mayoritaria}}{\# \text{ instancias en la clase minoritaria}}$$

Un banco de datos se considera desbalanceado, si  $IR > 1.5$ ; de lo contrario se dice que el banco de datos está balanceado.

Este es un problema que se presenta a menudo en situaciones comunes de la vida real.

## Desbalance de Clases

- Porque es importante atacar este problema:
  - **Pobre desempeño de los clasificadores:** regresión logística, maquinas de soporte vectorial y árboles de decisión.
  - Proceso de aprendizaje guiado por métricas globales presenta un BIAS hacia la clase mayoritaria.
  - Instancias de la clase minoritaria pueden ser ignorada.

## Desbalance de Clases

### 1. Métodos de muestreo: submuestreo

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.0	1.4	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
5.0	3.6	1.4	0.2	Setosa
4.4	2.9	1.4	0.2	Setosa
4.8	3.0	1.4	0.2	Setosa
5.0	3.3	1.4	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.5	4.9	1.3	Versicolor
6.9	3.1	4.9	1.5	Versicolor
6.2	3.4	5.4	2.3	Virginica
6.2	3.4	5.4	2.3	Virginica
6.2	3.4	5.4	2.3	Virginica



Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.0	1.4	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
4.8	3.0	1.4	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.5	4.9	1.3	Versicolor
6.9	3.1	4.9	1.5	Versicolor
6.2	3.4	5.4	2.3	Virginica
6.2	3.4	5.4	2.3	Virginica
6.2	3.4	5.4	2.3	Virginica



## Desbalance de Clases

### 1. Métodos de muestreo: sobremuestreo

Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.0	1.4	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
4.8	3.0	1.4	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
6.2	3.4	5.4	2.3	Virginica
6.2	3.4	5.4	2.3	Virginica



Largo de Sépalo	Ancho de Sépalo	Largo de Pétalo	Ancho de Pétalo	Clase
4.9	3.0	1.4	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
4.8	3.0	1.4	0.2	Setosa
6.9	3.1	4.9	1.5	Versicolor
5.7	3.5	4.9	1.3	Versicolor
6.9	3.1	4.9	1.5	Versicolor
6.2	3.4	5.4	2.3	Virginica
6.1	3.8	5.2	2.6	Virginica
6.0	3.1	5.8	2.1	Virginica



## Desbalance de Clases

3. Deshacerse de la clase minoritaria y tratar el problema con uno de detección de anomalías.

4. A nivel de algoritmo:

- Ajustar el peso de la clase.
- Ajustar el umbral de decisión.
- Modificar el algoritmo para que sea más sensitivo a clases raras.

# Referencias

- [1] **Leondes, C.T. (2018).** *Image Processing and Pattern Recognition*. California: Academic Press.
- [2] **Duda, R.O., Hart, P.E. & Stork, D.G. (2001).** *Pattern Classification*. 2<sup>nd</sup> edition. Wiley-Interscience.
- [3] **Marques de Sá, J:P. (2001).** *Pattern Recognition: Concepts, Methods and Applications*. Berlin: Springer-Verlag.
- [4] **Kuncheva, L. (2014).** *Combining Pattern Classifiers: Methods and Algorithms*. 2<sup>nd</sup> edition. USA: Wiley.
- [5] **Witten, I.H., Frank, E. & Hall, M.A. (2011).** *Data Mining: Practical Machine Learning Tools and Techniques*. 3<sup>rd</sup> edition. USA: Elsevier.
- [6] **Murty, N.M. & Devi, V.S. (2011).** *Pattern Recognition: An Algorithmic Approach*. Springer.
- [7] **Zaki, M.J. & Meira, W. (2014).** *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.
- [8] **Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017).** Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73, 220-239.