

Aprendizaje Automático No Supervisado

Alberto Barbado González

Tema 1 - Introducción

La pregunta del día

¿Cómo podemos identificar patrones y estructuras ocultas en conjuntos de datos sin etiquetas previas?

Encuesta previa

- ▶ ¿Cómo definirías de manera sencilla el UML?
- ▶ ¿Por qué crees que es necesario?
- ▶ Nombra una tarea que conozcas que se realiza con UML

En el día de hoy

- ▶ Definición de Aprendizaje Automático No Supervisado (UML)
- ▶ Aplicaciones de UML
- ▶ Flujo de experimentación con UML
- ▶ Preprocesamiento de datos

Definición

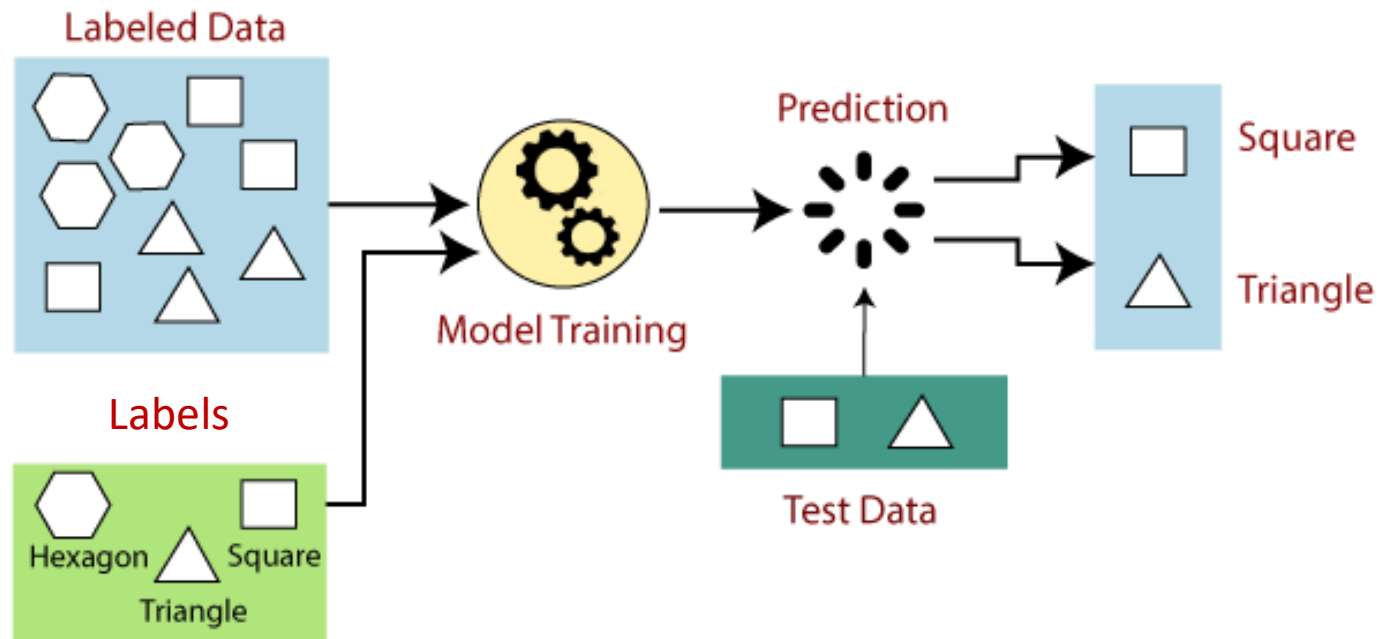
Unsupervised learning is a framework in machine learning where, in contrast to supervised learning, algorithms learn patterns exclusively from unlabeled data. ([Wikipedia](#))

Unsupervised learning in artificial intelligence is a type of machine learning that learns from data without human supervision. Unlike supervised learning, unsupervised machine learning models are given unlabeled data and allowed to discover patterns and insights without any explicit guidance or instruction ([Google](#))

El aprendizaje automático no supervisado se denomina así porque no hay un supervisor externo proporcionando retroalimentación sobre cómo debe ser clasificado un dato (Apuntes de la asignatura)

Ejemplos

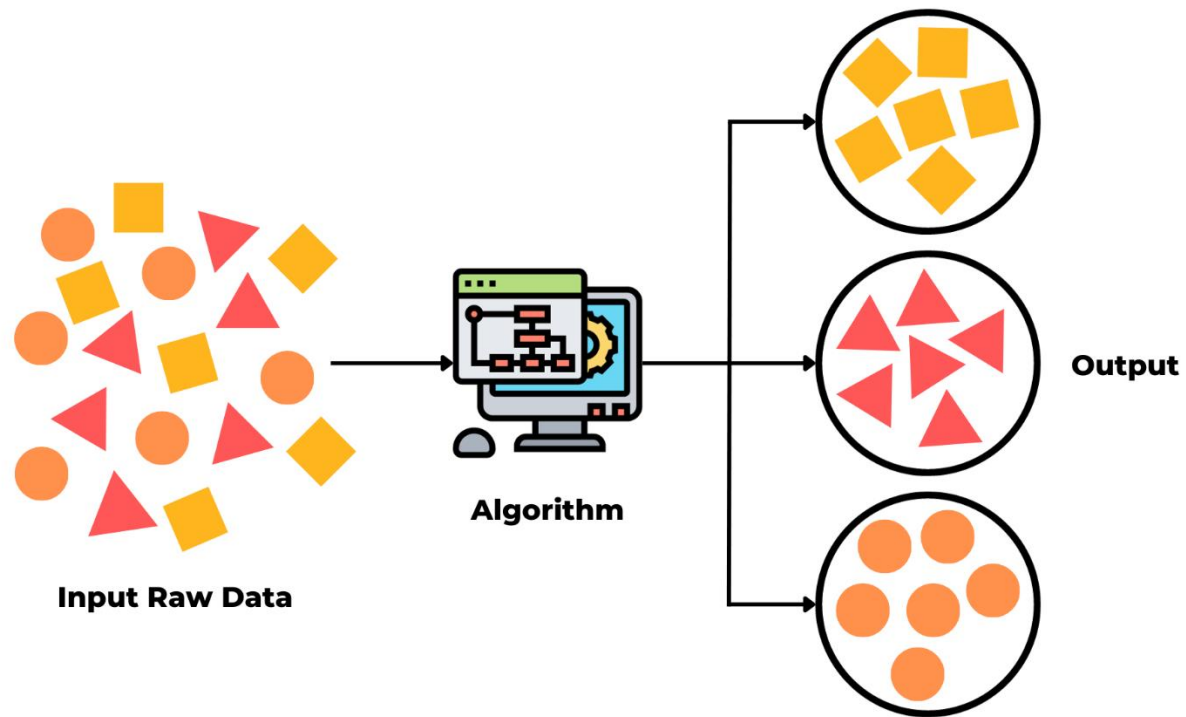
- ▶ Recordatorio: Aprendizaje Supervisado - Clasificación



<https://www.acte.in/supervised-learning-workflow-and-algorithms-article>

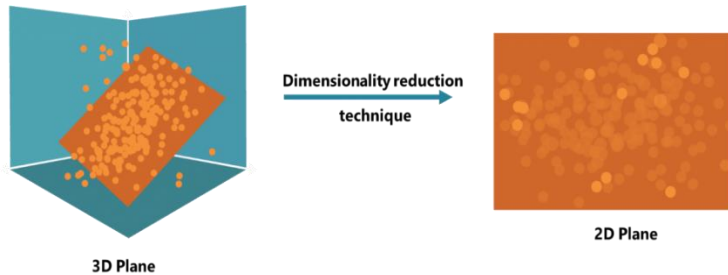
Ejemplos

- ▶ Aprendizaje No Supervisado - Clustering

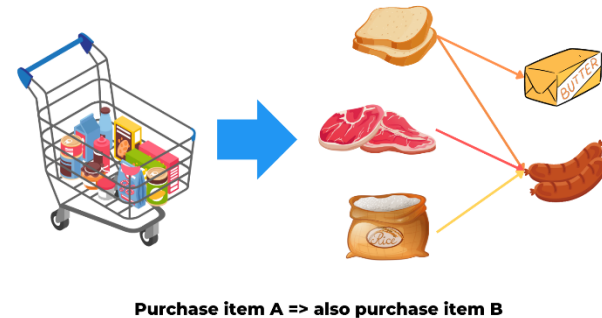


Ejemplos de UML

Reducción de dimensionalidad



Reglas de asociación



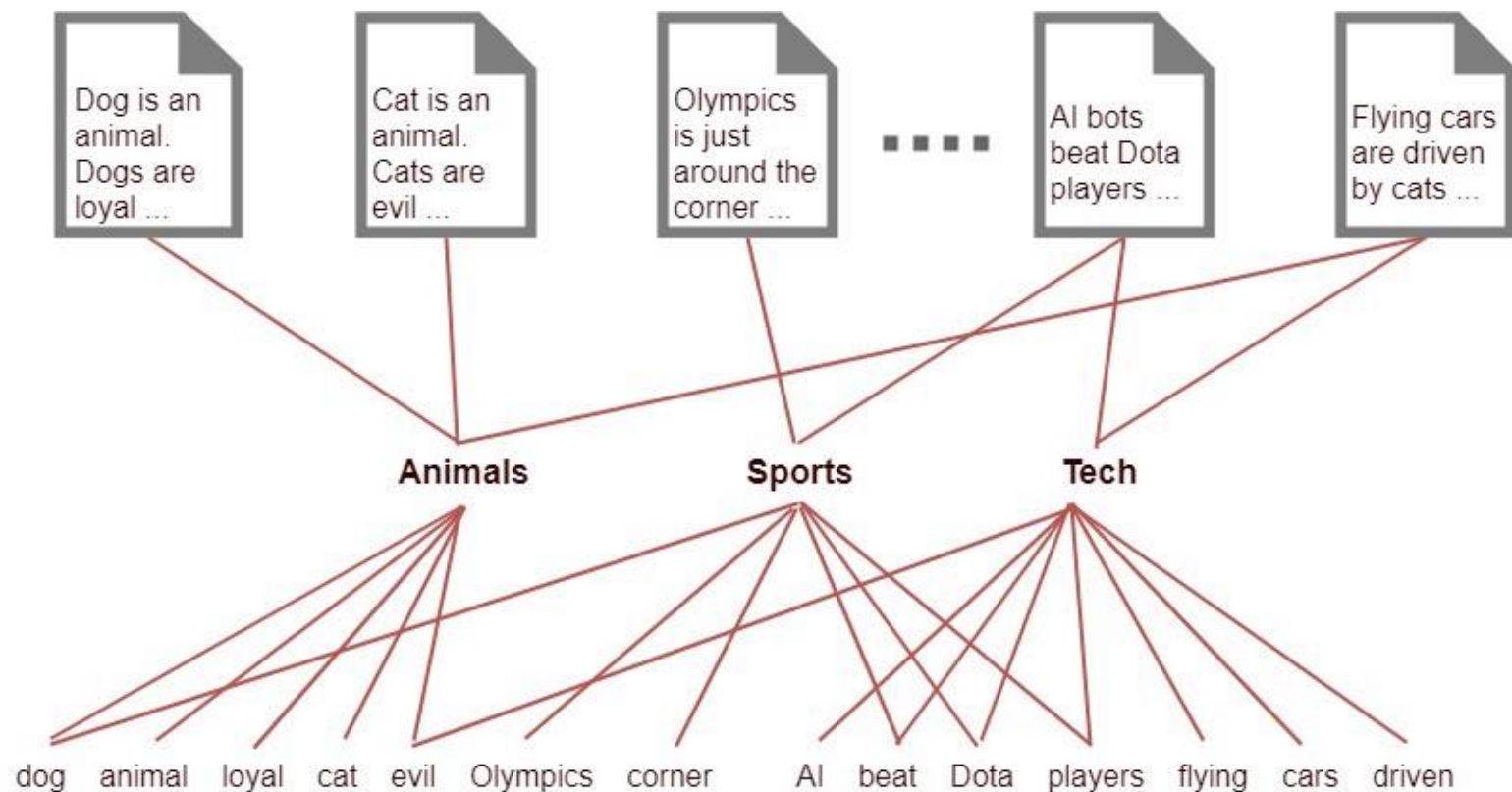
Detección de anomalías



<https://365datascience.com/tutorials/python-tutorials/principal-components-analysis/>
<https://ai-ml-analytics.com/introduction-to-machine-learning-blog-1/>

UML para distintos ámbitos: NLP

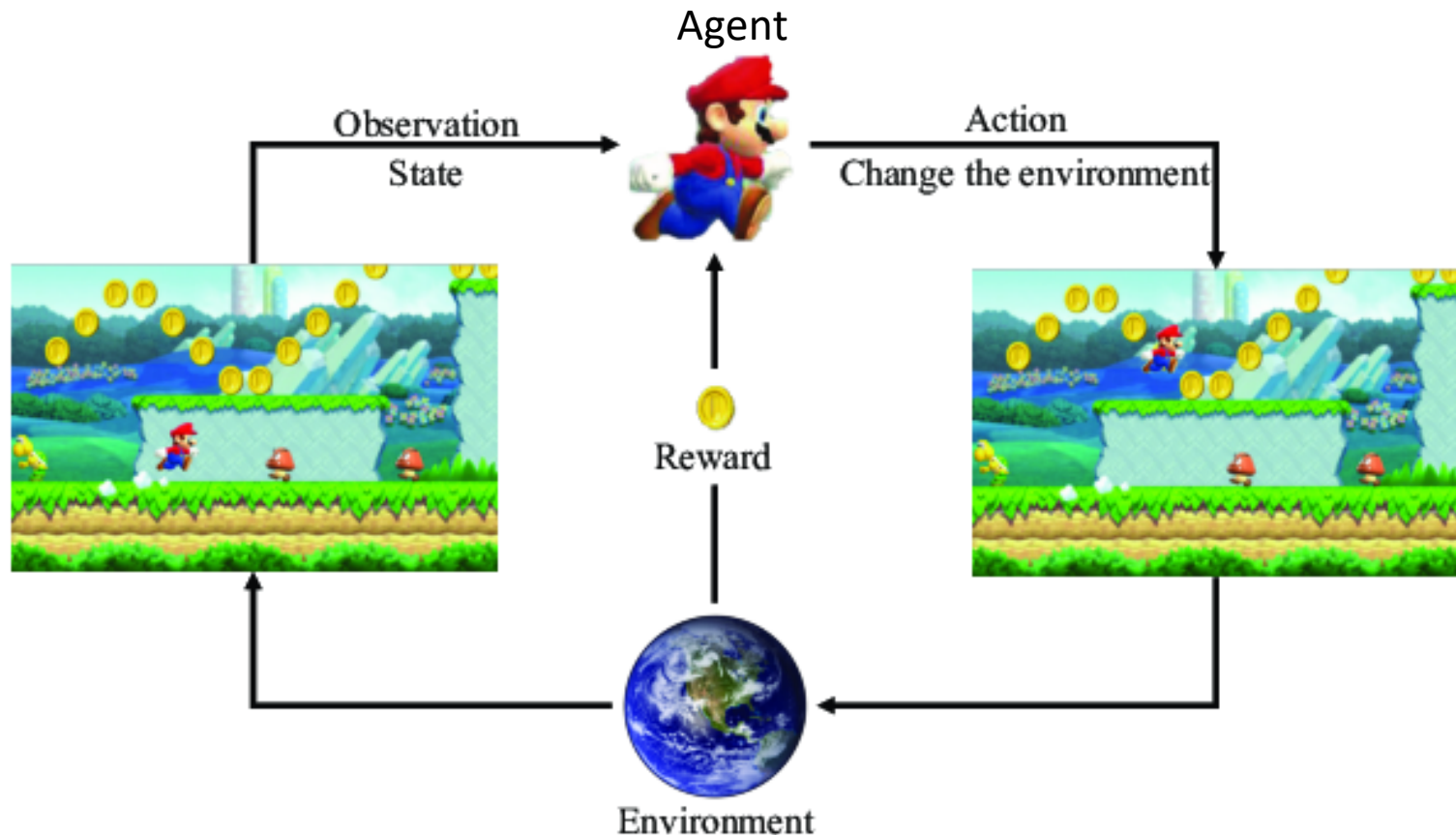
► Topic Modeling



<https://hyperskill.org/learn/step/27852>

ML que no es SML – Reinforcement Learning

- ▶ RL como otro caso que no es SML



A-TD3: An Adaptive Asynchronous Twin Delayed Deep Deterministic for Continuous Action Spaces

RL no es sólo para juegos...



- Gracias a técnicas de RL, se han podido desarrollar los LLMs tan eficientes que tenemos hoy en día.
- En particular, el caso de DeepSeek, donde, usando RL para desarrollar modelos con razonamiento que se “autocorrigen” antes de generar la respuesta final.

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

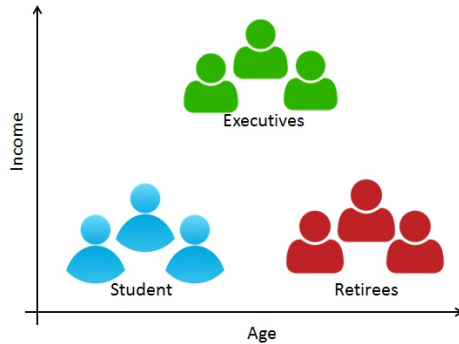


<https://arxiv.org/pdf/2501.12948>























<https://www.scmp.com/tech/big-tech/article/3301986/ai-agent-manus-partners-alibabas-qwen-develop-chinese-version>

Algunas aplicaciones a distintas industrias...

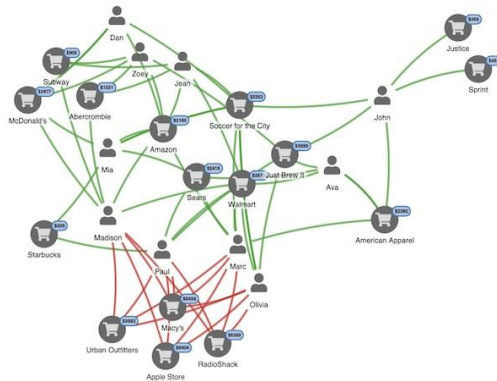
Segmentación de clientes



Ubicar productos en tienda

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

Detección de fraude en transacciones financieras

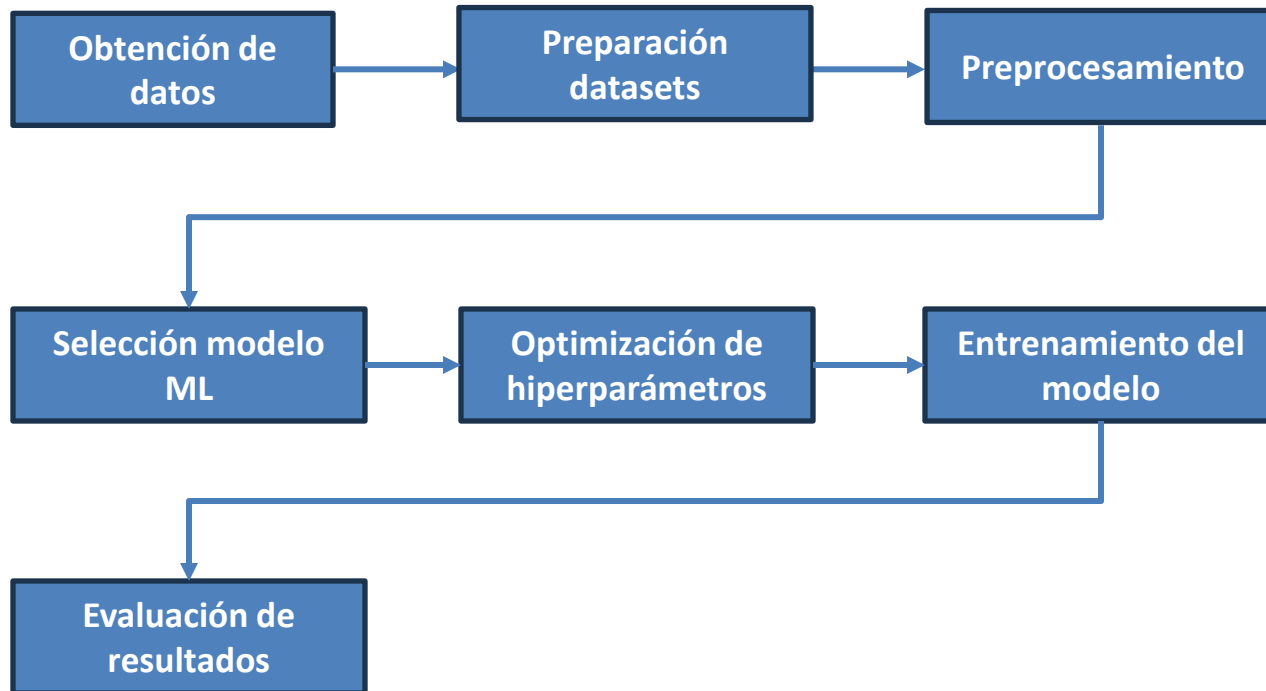


Personalizar contenido

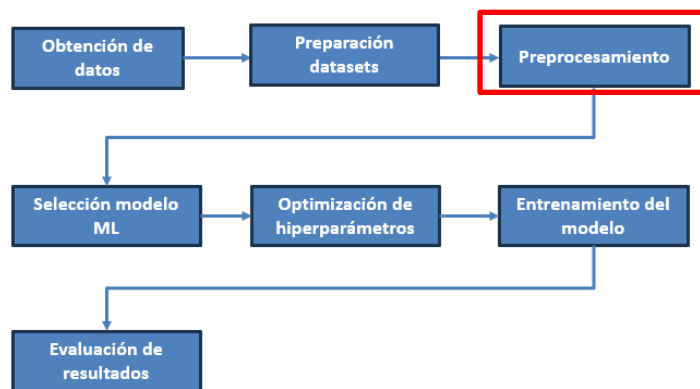


<https://rpubs.com/michelarrudala/801795>
<https://cambridge-intelligence.com/detect-credit-card-fraud-with-network-visualization/>
<https://www.kirenz.com/blog/posts/2020-05-14-r-association-rule-mining/>
<https://applemagazine.com/netflix-introduces-personalized-my-netflix-section-for-ios-users/59507>

Flujo de experimentación con UML



Preprocesamiento de datos



- Fase esencial para **convertir información en datos** procesables por el **computador**.
- Generalmente, los datos deben ser **numéricos**; textos e imágenes requieren transformación.
- La **calidad del procesamiento** es clave para asegurar resultados precisos.

<https://www.linkedin.com/pulse/garbage-out-chanthoeun-chiv/>

Preprocesamiento - técnicas

Las habituales...

- ▶ Tratamiento de nulos
- ▶ Transformación de datos
- ▶ Discretización
- ▶ Estandarización
- ▶ Codificación de datos categóricos
- ▶ Tratamiento de valores atípicos (outliers)

Preprocesamiento - técnicas

Otras más específicas a cada situación...

- ▶ Reducción de dimensionalidad
 - ▶ Selección de variables
 - ▶ Extracción de variables
- ▶ Preprocesamiento de imágenes
- ▶ Preprocesamiento de texto
 - ▶ Tokenización
 - ▶ Eliminación de stopwords
 - ▶ Stemming / lematización
 - ▶ NER
 - ▶ ...
- ▶ *Resampling techniques / Data Augmentation*

Preprocesamiento - técnicas

Otras más específicas a cada situación...

- ▶ Reducción de dimensionalidad

- ▶ Selección de variables
- ▶ Extracción de variables



Lo veremos en la asignatura

- ▶ Preprocesamiento de imágenes



Visión Artificial

- ▶ Preprocesamiento de texto

- ▶ Tokenización
- ▶ Eliminación de stopwords
- ▶ Stemming / lematización
- ▶ NER
- ▶ ...



Procesamiento de Lenguaje Natural

- ▶ *Resampling techniques / Data Augmentation*

Tratamiento de nulos

Objetivo: Tratamiento de valores nulos o incompletos

Age	Salary	Department
25	50000	HR
30	NaN	Finance
NaN	60000	IT

Algunas técnicas

- Eliminar registros
- Reemplazar con media, mediana o moda
- Reemplazar con valor por defecto
- Reemplazar por UNK
- Predecirlo con un modelo de ML!

Ejemplo: Reemplazar por media

Age	Salary	Department
25	50000	HR
30	55000	Finance
27.5	60000	IT

Tratamiento de nulos

Age	Salary	Department
25	50000	HR
30	-1	Finance
99999	60000	IT

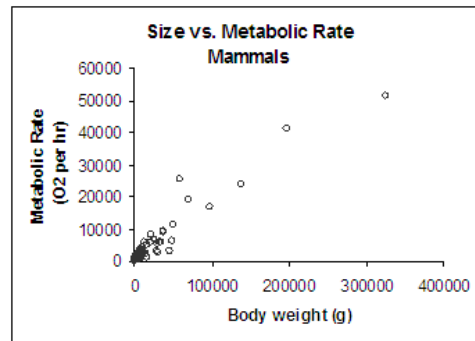
Nota: Los nulos a veces no se representan como valores "vacíos"

Ejemplo: Reemplazar por media

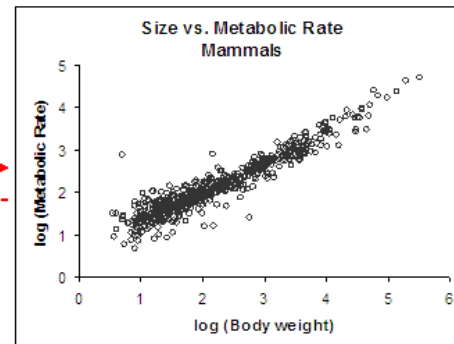
Age	Salary	Department
25	50000	HR
30	55000	Finance
27.5	60000	IT

Transformación de datos

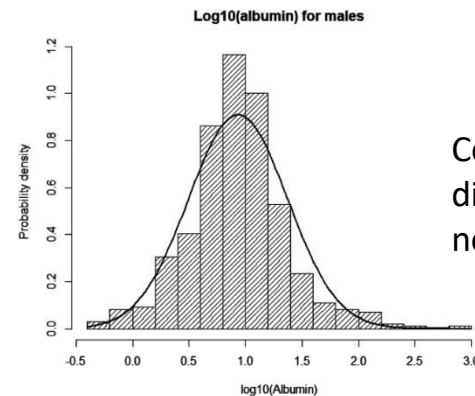
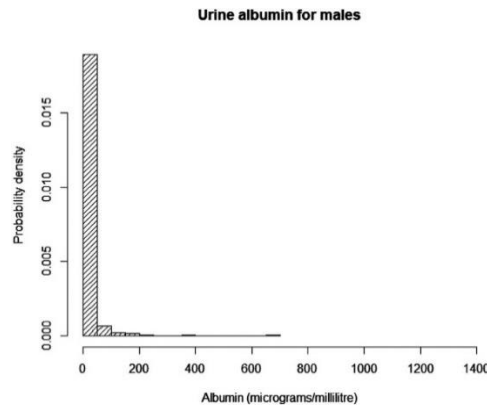
Objetivo: Reducir asimetría, convertir relaciones no lineales en lineales y mejorar la predicción.



Log
→
Transform



Ayuda a tener relaciones lineales



Contribuye a tener distribuciones normales / gaussianas

https://mathbench.umd.edu/modules/misc_scaling/page07.htm
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9036143/>

Transformación de datos

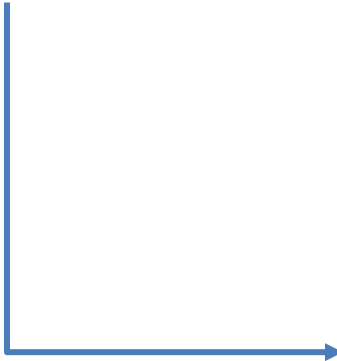
Objetivo: Reducir asimetría, convertir relaciones no lineales en lineales y mejorar la predicción.

Age	Salary	Department
25	50000	HR
30	55000	Finance
27	4500000	IT

Algunas técnicas

- Transformación logarítmica
- Transformación polinómica

Ejemplo: Transformación logarítmica

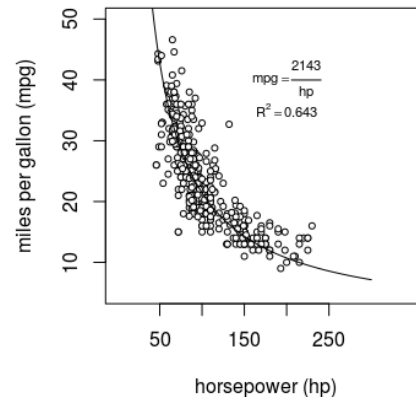
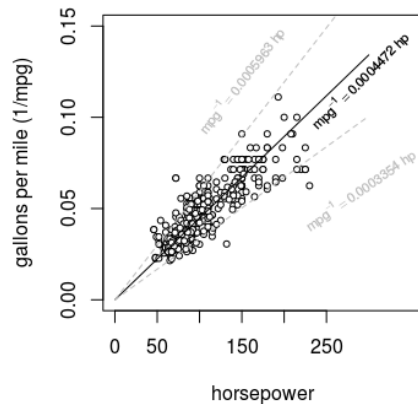


Age	Ln(Salary)	Department
25	10.82	HR
30	10.91	Finance
27	15.31	IT

Transformación de datos

Objetivo: Reducir asimetría, convertir relaciones no lineales en lineales y mejorar la predicción.

Ejemplo de aplicación



Ejemplos de transformaciones polinómicas

Transformation	$f(x)$ Notation	Examples
Horizontal Translation Graph shifts left or right.	$f(x - h)$	$g(x) = (x - 5)^4$ 5 units right $g(x) = (x + 2)^4$ 2 units left
Vertical Translation Graph shifts up or down.	$f(x) + k$	$g(x) = x^4 + 1$ 1 unit up $g(x) = x^4 - 4$ 4 units down
Reflection Graph flips over x - or y -axis.	$f(-x)$ $-f(x)$	$g(x) = (-x)^4 = x^4$ over y -axis $g(x) = -x^4$ over x -axis
Horizontal Stretch or Shrink Graph stretches away from or shrinks toward y -axis.	$f(ax)$	$g(x) = (2x)^4$ shrink by a factor of $\frac{1}{2}$ $g(x) = \left(\frac{1}{2}x\right)^4$ stretch by a factor of 2
Vertical Stretch or Shrink Graph stretches away from or shrinks toward x -axis.	$a \cdot f(x)$	$g(x) = 8x^4$ stretch by a factor of 8 $g(x) = \frac{1}{4}x^4$ shrink by a factor of $\frac{1}{4}$

https://static.bigideasmath.com/protected/content/pe/hs/sections/alg2_pe_04_07.pdf

<https://stats.stackexchange.com/questions/568417/why-would-you-perform-transformations-over-polynomial-regression>

Discretización de datos

Objetivo: Convertir los valores de una variable continua en valores discretos, generando intervalos o *bins*.

Age
25
30
27
40

Ejemplo: Uso de cuantiles ($q=3$)

Age	Age_Bin
25	0
30	1
27	0
40	2

Algunas técnicas

- Uso de cuantiles
- Árboles de Decisión
- Clustering

Estandarización de datos

Objetivo: Asegurar que los datos están en una misma escala. Esto es fundamental para usarlos en algoritmos basados en distancias (e.j., NN)

Age	Salary	Department
25	50000	HR
30	55000	Finance
27	80000	IT



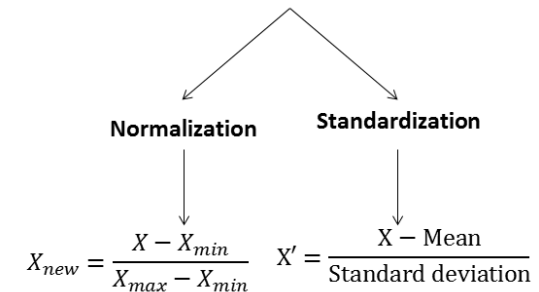
Ejemplo: Normalización

Age	Salary	Department
0	0	HR
1	0.167	Finance
0.625	1	IT

Algunas técnicas

- Normalización
- Estandarización

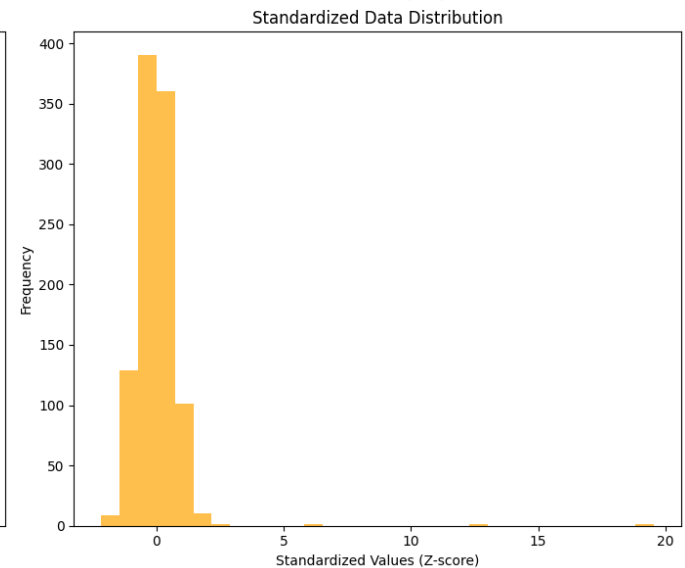
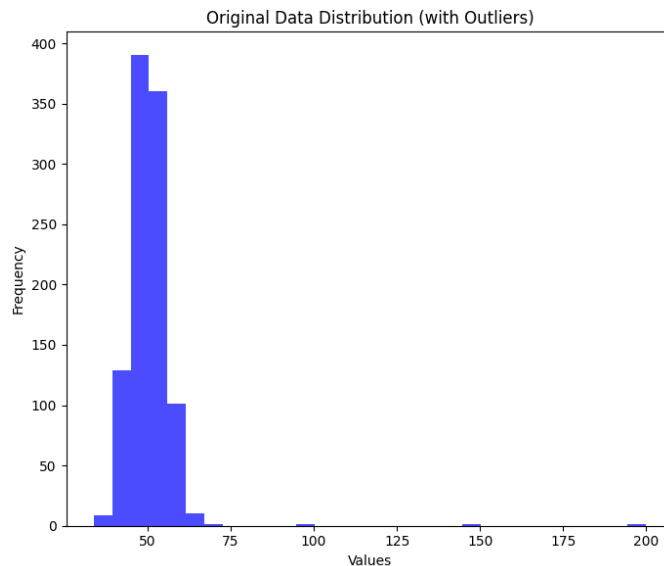
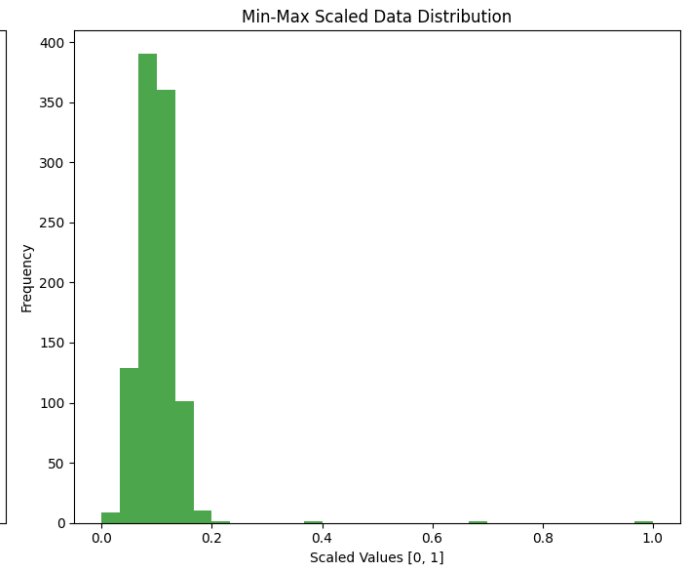
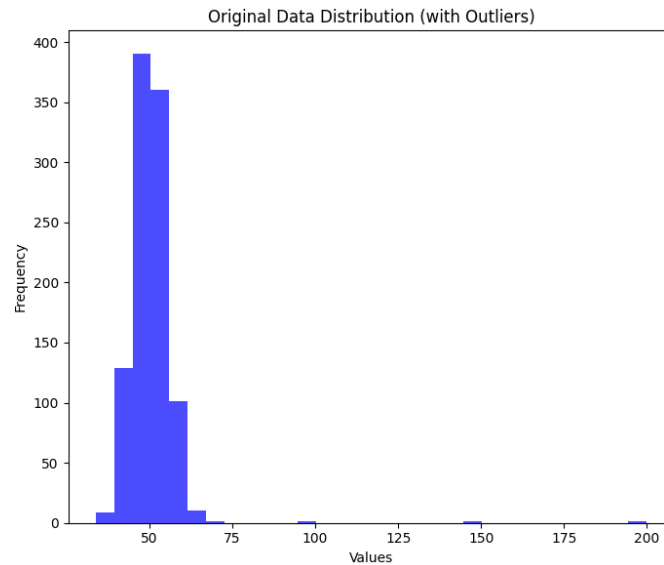
Feature scaling



https://medium.com/@brijesh_soni/feature-engineering-101-7cb68d293551

Estandarización de datos

Influencia de
los valores
atípicos




Codificación de valores categóricos

Objetivo: Convertir los datos textuales en formato numérico

Degree	Department
Graduate	HR
MSc	Finance
PhD	IT

Algunas técnicas

- One-hot encoding
- Label encoding
- Codificación binaria



Contract	Department_HR	Department_Finance	Department_IT
0	1	0	0
1	0	1	0
2	0	0	1

Label encoding

One-hot encoding

Tratamiento de outliers

Objetivo: Detectar y tratar *outliers* que pueden perjudicar al rendimiento del modelo.

Serie = [10, 12, 14, 15, 16, 18, 19, 21, 100]

```
Q1: 14.0
Q3: 19.0
IQR: 5.0
IQR Outliers (outside [ 6.5 , 26.5 ]): [100]
```

Algunas técnicas

- Z-score
- IQR

Fórmulas

$$\text{Z-Score Formula: } Z = \frac{X - \mu}{\sigma}$$

$$\text{IQR} = Q3 - Q1$$

E.j. Umbrales

$$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$$

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$$

$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$$

$$\text{Lower bound} = Q1 - 1.5 \times \text{IQR}$$

$$\text{Upper bound} = Q3 + 1.5 \times \text{IQR}$$

En resumen

- ▶ UL como enfoque para encontrar patrones en datos cuando no se dispone de un target específico
- ▶ UL como un conjunto de algoritmos útiles para distintos casos de uso: perfilado clientes, detección de fraude, ubicación de productos en tiendas, personalización de contenido...
- ▶ Preprocesado como etapa fundamental inicial de todos los flujos de experimentación y trabajo de ML (SML y UML)

En la próxima semana

¿Cómo podemos agrupar un conjunto de datos de clientes de una tienda en diferentes segmentos sin conocer previamente su comportamiento de compras?

UNIVERSIDAD
INTERNACIONAL
DE LA RIOJA

unir

www.unir.net