PREPROCESADO DE DATOS

Pablo Pérez Núñez

Índice

Índice

Introducción

Preprocesado de datos:

- Tratar los datos buscando **mejorar su calidad** con el fin de aumentar la eficacia de los modelos de aprendizaje automático.

Consta de:

- **Análisis**
- **Limpieza**
- **Normalización**
- **Estandarización**

Introducción

Encargo de programa:

- Quiero un programa que dados los datos académicos de un alumno me calcule si la nota final es aprobado o no.
- Quiero hacerlo como se hizo en otros años, pero no me dijeron cuál era el criterio.
- Tengo datos de años pasados:

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
…	…	…	…	…	…
A375	100375	True	None	0.9	4.9

Tabla 1: *Ejemplo de datos.*

Introducción

Encargo de programa: Ciclo del aprendizaje automático

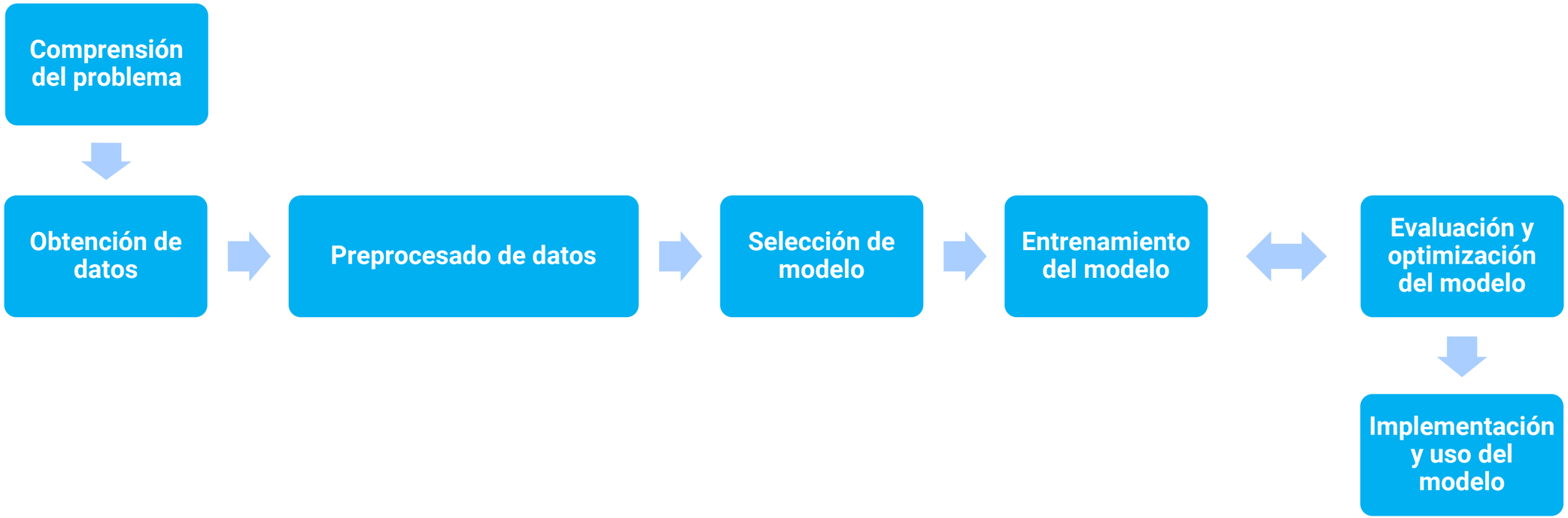

Figura 1: *Procedimiento a seguir a la hora de resolver un problema de aprendizaje automático.*

Introducción

Encargo de programa: Ciclo del aprendizaje automático

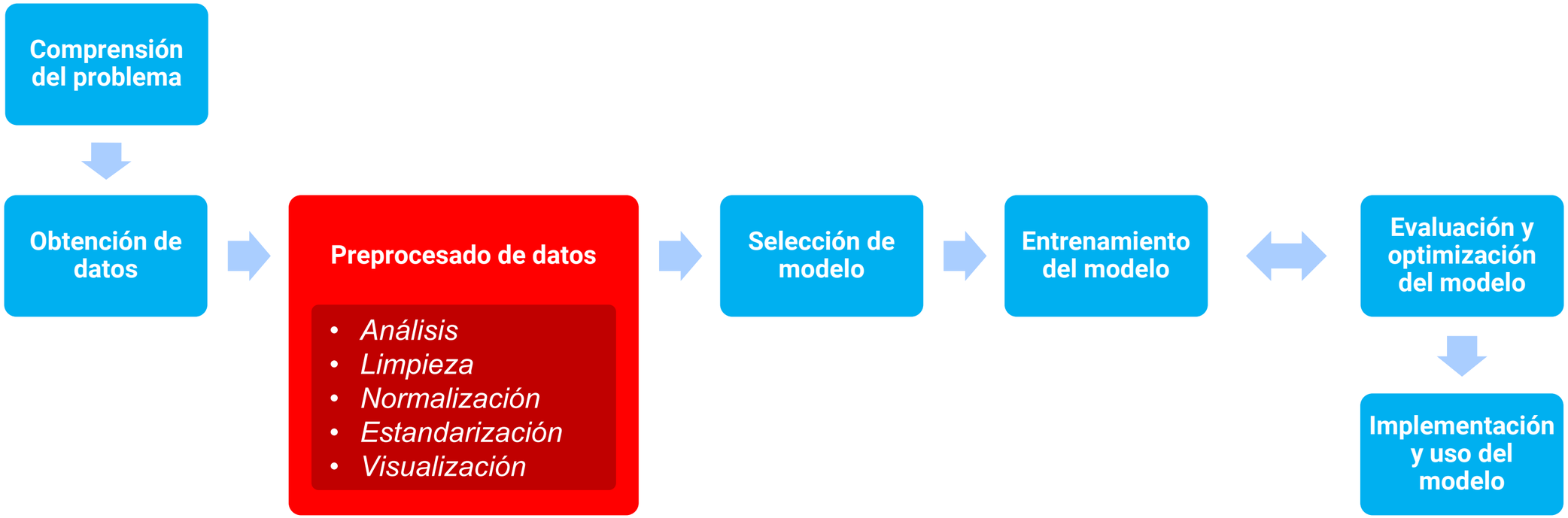

Figura 1: *Procedimiento a seguir a la hora de resolver un problema de aprendizaje automático.*

Introducción

Encargo de programa: Ciclo del aprendizaje automático

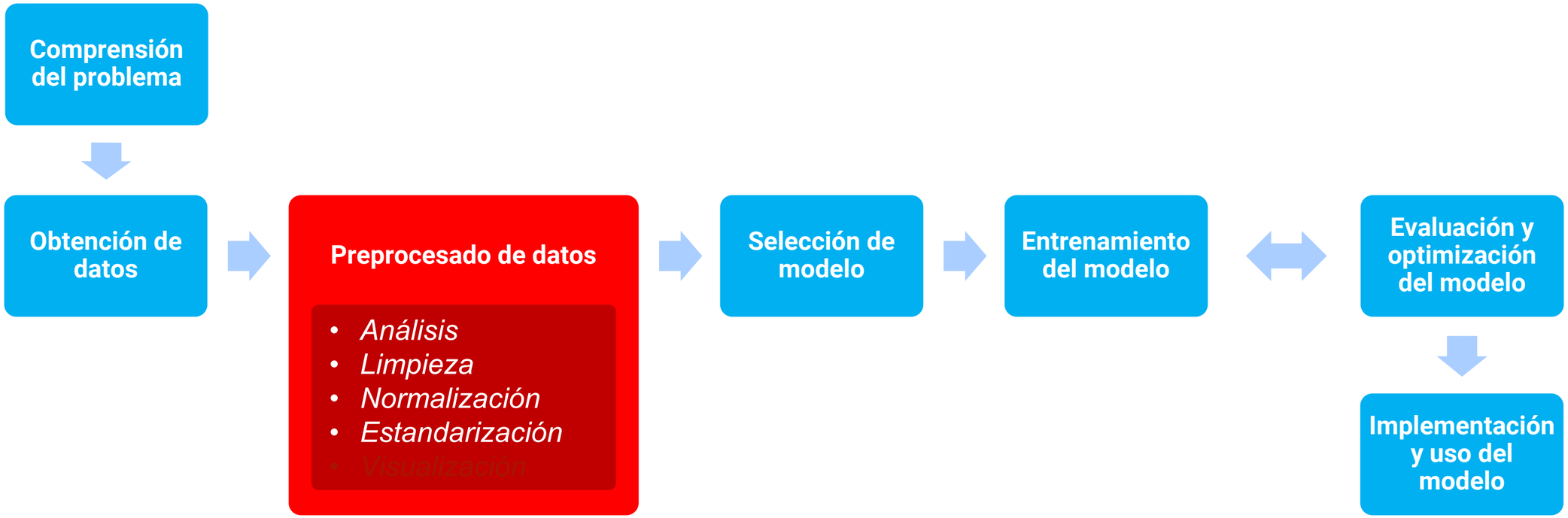

Figura 1: *Procedimiento a seguir a la hora de resolver un problema de aprendizaje automático.*

Índice

Análisis y limpieza

Análisis:

- **Tipo de datos:**
  - Tradicionales:
    - Numéricos.
    - Textuales.

Análisis y limpieza

Análisis:

- **Tipo de datos:**
  - Multimedia:
    - Imagen
    - Video
    - Audio

Habitualmente los conjuntos son **multimodales**, es decir, mezclan varios de estos tipos.

Análisis y limpieza

Análisis:

- Formato de los archivos:

Formato	Descripción
.txt	Texto plano.
.csv	Comma Separated Values.
.xlsx	Microsoft Excel spreadsheet files.
.json	JavaScript Object Notation.
.pkl	Pickle (Python).
.arff	Attribute-Relation File Format (Weka).
.sqlite	Base de datos SQLite.

Tabla 2: *Extensiones comunes de los archivos de datos tradicionales.*

Análisis y limpieza

Análisis:

- Formato de los archivos:

Formato	Descripción	Compresión
.jpeg	Joint Photographic Experts Group.	Si
.png	Portable Network Graphics.	Si/No
.raw	Imágenes en “crudo”.	No
.nifti	Neuroimaging Informatics Technology Initiative.	Si/No
.wav	Waveform Audio File Format.	No
.mp3	MPEG-2 Audio Layer III.	Si

Tabla 3: *Extensiones comunes de los archivos de datos multimedia.*

Análisis y limpieza

Análisis:

- **Relación entre datos:**
  - El conjunto puede estar formado por un solo fichero o por múltiples interconectados.

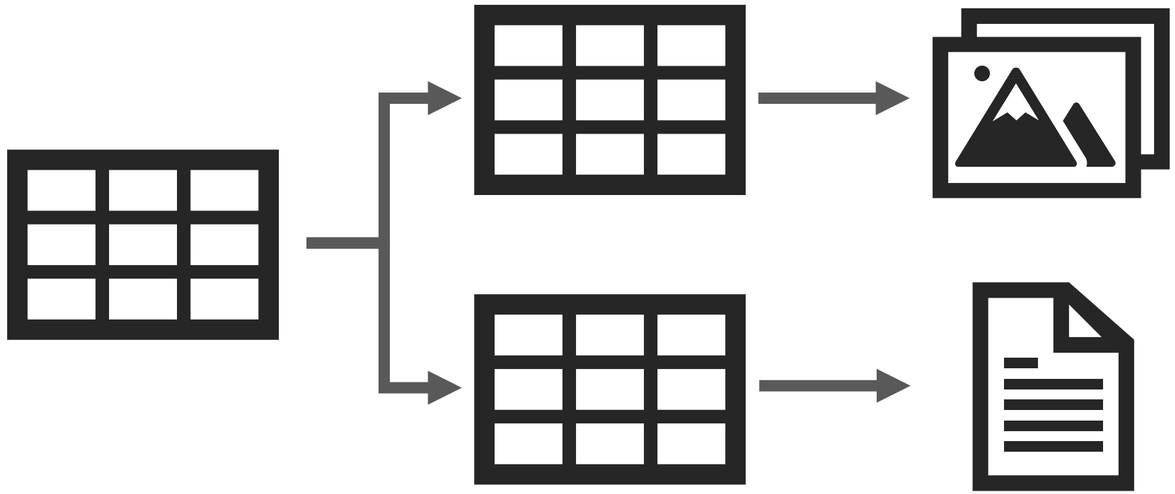

Figura 2: *Estructura de datos de ejemplo con tres tablas, imágenes y documentos.*

Análisis y limpieza

Análisis:

- **Volumen de datos:**
  - ¿Podemos manejarlo en nuestro ordenador personal?
  - ¿Se puede cargar completamente en memoria?
  - ¿Necesitamos dividirlo en subconjuntos?
  - ¿Tenemos que instalar un sistema de gestión de bases de datos?
  - ¿Son datos en tiempo real?

Análisis y limpieza

Análisis:

- **Suficiencia de los datos:**
  - Determinar si los datos disponibles son **suficientes para aprender el problema.**
  - A **mayor complejidad**, **mayor cantidad y diversidad** de datos requeridas.
- **Tamaños de conjuntos típicos:**

Nombre	Descripción
Iris	150 muestras con 4 características
CIFAR-100	60.000 imágenes de 32x32 pixels distribuidas en 100 categorías
COCO	330.000 imágenes con 80 categorías de objetos.
ImageNet	Más de 14 millones de imágenes en 1000 categorías.
Wikipedia	Millones de páginas web y terabytes de texto.

Tabla 4: *Tamaño de conjuntos de datos conocidos.*

Análisis y limpieza

Limpieza:

- **Errores y valores ausentes:**
  - Es muy común encontrar valores ausentes o columnas mal calculadas.
  - En muchos casos se pueden recuperar los valores, en otros hay que eliminarlos.

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
…	…	…	…	…	…
100374	False	7.6	0.4	10.0	
A375	100375	True	None	0.9	4.9

Tabla 5: *Ejemplo de datos.*

Análisis y limpieza

Limpieza:

- **Errores y valores ausentes:**
  - Es muy común encontrar valores ausentes o columnas mal calculadas.
  - En muchos casos se pueden recuperar los valores, en otros hay que eliminarlos.

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
...	...	...	...	...	...
100374	False	7.6	0.4	10.0	
A375	100375	True	None	0.9	4.9

Tabla 5: *Ejemplo de datos.*

Análisis y limpieza

Limpieza:

- **Errores y valores ausentes:**
  - Es muy común encontrar valores ausentes o columnas mal calculadas.
  - En muchos casos se pueden recuperar los valores, en otros hay que eliminarlos.

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
…	…	…	…	…	…
100374	False	7.6	0.4	10.0	
A375	100375	True	None	0.9	4.9

Tabla 5: *Ejemplo de datos.*

Análisis y limpieza

Limpieza:

- **Errores y valores ausentes:**
  - Es muy común encontrar valores ausentes o columnas mal calculadas.
  - En muchos casos se pueden recuperar los valores, en otros hay que eliminarlos.

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
…	…	…	…	…	…
A374	100374	False	7.6	0.4	10.0
A375	100375	True	4.0	0.9	4.9

Tabla 6: *Ejemplo de datos sin errores.*

Análisis y limpieza

Limpieza:

- **Variables irrelevantes:**
  - Muchas veces las columnas son prescindibles en la tarea a resolver.

Nombre	UO	Aprobado?	Teoría	Prácticas	Suma notas
A001	100001	False	8	0.04	8.04
A002	100002	False	10	0.35	10.35
…	…	…	…	…	…
A374	100374	False	7.6	0.4	10.0
A375	100375	True	4.0	0.9	4.9

Tabla 6: *Ejemplo de datos sin errores.*

Análisis y limpieza

Limpieza:

- **Variables irrelevantes:**
  - Muchas veces las columnas son prescindibles en la tarea a resolver.

~~Nombre~~	~~UO~~	Aprobado?	Teoría	Prácticas	Suma notas
~~A001~~	~~100001~~	False	8	0.04	8.04
~~A002~~	~~100002~~	False	10	0.35	10.35
~~...~~	~~...~~	...	...	...	...
~~A374~~	~~100374~~	False	7.6	0.4	10.0
~~A375~~	~~100375~~	True	4.0	0.9	4.9

Tabla 6: *Ejemplo de datos sin errores.*

Análisis y limpieza

Limpieza:

- **Variables irrelevantes:**
  - Muchas veces las columnas son prescindibles en la tarea a resolver.

Aprobado?	Teoría	Prácticas	Suma notas
False	8	0.04	8.04
False	10	0.35	10.35
…	…	…	…
False	7.6	0.4	10.0
True	4.0	0.9	4.9

Tabla 7: *Ejemplo de datos sin errores ni variables irrelevantes.*

Análisis y limpieza

Limpieza:

- **Relación entre variables:**
  - Analizar las columnas en busca de correlaciones entre ellas.
  - **Métodos:** Information Gain o Correlation-based feature selection.

Aprobado?	Teoría	Prácticas	Suma notas
False	8	0.04	8.04
False	10	0.35	10.35
…	…	…	…
False	7.6	0.4	10.0
True	4.0	0.9	4.9

Tabla 7: *Ejemplo de datos sin errores ni variables irrelevantes.*

Análisis y limpieza

Limpieza:

- **Relación entre variables:**
  - Analizar las columnas en busca de correlaciones entre ellas.
  - **Métodos:** Information Gain o Correlation-based feature selection.

	Teoría	Prácticas	Suma notas
Teoría	100%	-1%	73%
Prácticas	-1%	100%	68%
Suma notas	73%	68%	100%

Figura 3: *Matriz de correlación de las columnas numéricas del conjunto.*

Análisis y limpieza

Limpieza:

- **Relación entre variables:**
  - Analizar las columnas en busca de correlaciones entre ellas.
  - **Métodos:** Information Gain o Correlation-based feature selection.

	Teoría	Prácticas	Suma notas
Teoría	100%	-1%	73%
Prácticas	-1%	100%	68%
Suma notas	73%	68%	100%

Figura 3: *Matriz de correlación de las columnas numéricas del conjunto.*

Análisis y limpieza

Limpieza:

- **Relación entre variables:**
  - Analizar las columnas en busca de correlaciones entre ellas.
  - **Métodos:** Information Gain o Correlation-based feature selection.

Aprobado?	Teoría	Prácticas	~~Suma notas~~
False	8	0.04	~~8.04~~
False	10	0.35	~~10.35~~
…	…	…	~~…~~
False	7.6	0.4	~~10.0~~
True	4.0	0.9	~~4.9~~

Tabla 7: *Ejemplo de datos sin errores ni variables irrelevantes.*

Análisis y limpieza

Limpieza:

- **Relación entre variables:**
  - Analizar las columnas en busca de correlaciones entre ellas.
  - **Métodos:** Information Gain o Correlation-based feature selection.

Aprobado?	Teoría	Prácticas
False	8	0.04
False	10	0.35
…	…	…
False	7.6	0.4
True	4.0	0.9

Tabla 8: *Ejemplo de datos sin errores, variables irrelevantes ni redundancias.*

Índice

Estandarización y normalización

Objetivo:

- Igualar columnas con valores en diferentes escalas o distribuciones.
- Así, los algoritmos tratarán todas las columnas por igual y aprenderán mejor.

Aprobado?	Teoría	Prácticas
False	8.00	0.04
False	10.00	0.35
…	…	…
False	7.60	0.40
True	4.00	0.90

Media	6.60	0.45
Desviación	3.91	0.42
Máximo	10.00	1.00
Mínimo	0.00	0.00

Tabla 9: *Estadísticas básicas del conjunto resultante.*

Estandarización y normalización

Estandarización:

- Escalar los datos buscando **media 0 y desviación 1.** Distribución gaussiana.

$$z = \frac{(x - \bar{x})}{s}$$

Aprobado?	Teoría	Prácticas
False	0.36	-0.97
False	0.87	-0.23
…	…	…
False	0.26	-0.11
True	-0.67	1.07

Media	**0.00**	**0.00**
Desviación	**1.00**	**1.00**
Máximo	0.87	1.31
Mínimo	-1.69	-1.06

Tabla 10: *Estandarización y estadísticas del conjunto resultante.*

Estandarización y normalización

Normalización:

- Escalar los datos **entre 0 y 1**.

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Aprobado?	Teoría	Prácticas
False	0.80	0.04
False	1.00	0.35
…	…	…
False	0.76	0.40
True	0.40	0.90

Media	0.66	0.45
Desviación	0.39	0.42
Máximo	**1.00**	**1.00**
Mínimo	**0.00**	**0.00**

Tabla 11: *Normalización y estadísticas del conjunto resultante.*

Estandarización y normalización

Normalización:

- Escalar los datos **entre -1 y 1**.

$$z = 2 * \left(\frac{x - min(x)}{max(x) - min(x)} \right) - 1$$

Aprobado?	Teoría	Prácticas
False	0.60	-0.92
False	1.00	-0.30
…	…	…
False	0.52	-0.20
True	-0.20	0.80

Media	0.32	-0.10
Desviación	0.78	0.84
Máximo	**1.00**	**1.00**
Mínimo	**-1.00**	**-1.00**

Tabla 12: *Normalización y estadísticas del conjunto resultante.*

Índice

Herramientas

- **Matemáticas:**
  - Matlab
  - R
- **Informáticas:**
  - Python:
    - Pandas
    - scikit-learn
- **Otras:**
  - Excel
  - PowerBi