

# Método KDD: Preprocesamiento

Daniel Marín López

Guadalupe Luna Velázquez

Víctor González Gómez





# Índice

1

## Introducción

Definición de preprocesamiento, ruido...

2

## Limpieza de los datos

Identificar y corregir errores, inconsistencias e inexactitudes.

3

## Valores desconocidos

Datos que deberían estar presentes, pero no se registraron.

4

## Outliers e inliers

Análisis de datos extremos y comunes: la distinción entre outliers e inliers.





5

## Valores Prototípicos y por defecto

Definición de valores prototípicos y por defecto en los datos.

6

## Valores erróneos

Son datos considerados incorrectos, inconsistentes o fuera de lugar dentro de un conjunto de datos.

7

## Agregación de datos

Reducir la cantidad de datos, enfocándose en la información más relevante.

# 1. Introducción: Preproceso

Es la fase en la que se limpia y prepara el conjunto de datos antes de su análisis o modelado.

Algunas cosas que se hacen en el preprocesamiento son:

- Eliminación de ruido
- Tratamiento de campos vacíos
- Outliers e inliers



## PREPROCESO





## 2. Limpieza de los datos

La limpieza de datos en el método KDD tiene el objetivo principal es mejorar la calidad de los datos para que las fases posteriores de minería de datos puedan generar resultados más fiables y significativos.

En esencia, la limpieza de datos consiste en identificar y corregir errores, inconsistencias e inexactitudes que puedan existir en el conjunto de datos seleccionado.



# Tareas de la limpieza de datos

## Manejar valores faltantes

Se debe tomar una decisión sobre los datos faltantes.

## Resolución de inconsistencias

Unificar diferentes formas de representar un mismo dato.



## Identificar y tratar el ruido

El ruido se describe como los datos que son erróneos o atípicos que no reflejan la realidad

## Corregir errores e inconsistencias

Aquí se buscan y tratan los errores tipográficos, valores fuera de rango, etc.

## Eliminar duplicados

Buscar y eliminar datos idénticos o redundantes para que no afecten en los resultados finales.

# Mejoras de la limpieza de datos

## Mejorar la precisión

Datos más limpios, mejor precisión en los modelos y sistemas.

## Reducir el riesgo de obtener conclusiones erróneas

Si minimizamos los datos defectuosos, evitaremos detectar patrones falsos y errores.

## Optimizar el rendimiento de los algoritmos

Datos de buena calidad, mayor rendimiento en nuestros sistemas.

## Facilitar la interpretación de los resultados

Datos más limpios, mejor comprensión y fiabilidad en los patrones.

# 3. Valores desconocidos

Los “missings values” son datos que deberían estar presentes, pero no se registraron por alguna razón conocida.

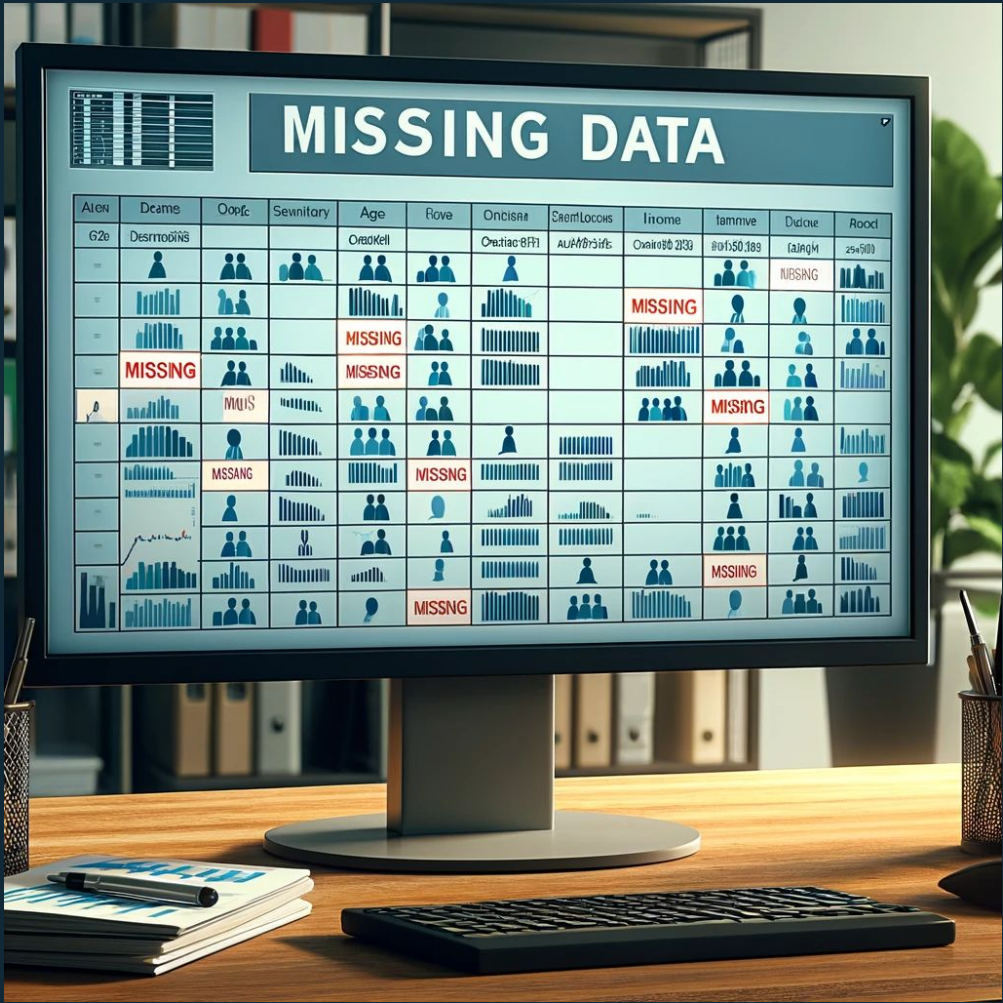
Missing values										
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q





Los “system missing values” son datos que no existen porque el sistema no los puede generar o procesar, y muchas veces están representados por un valor especial como “NaN” o simplemente no aparecen.

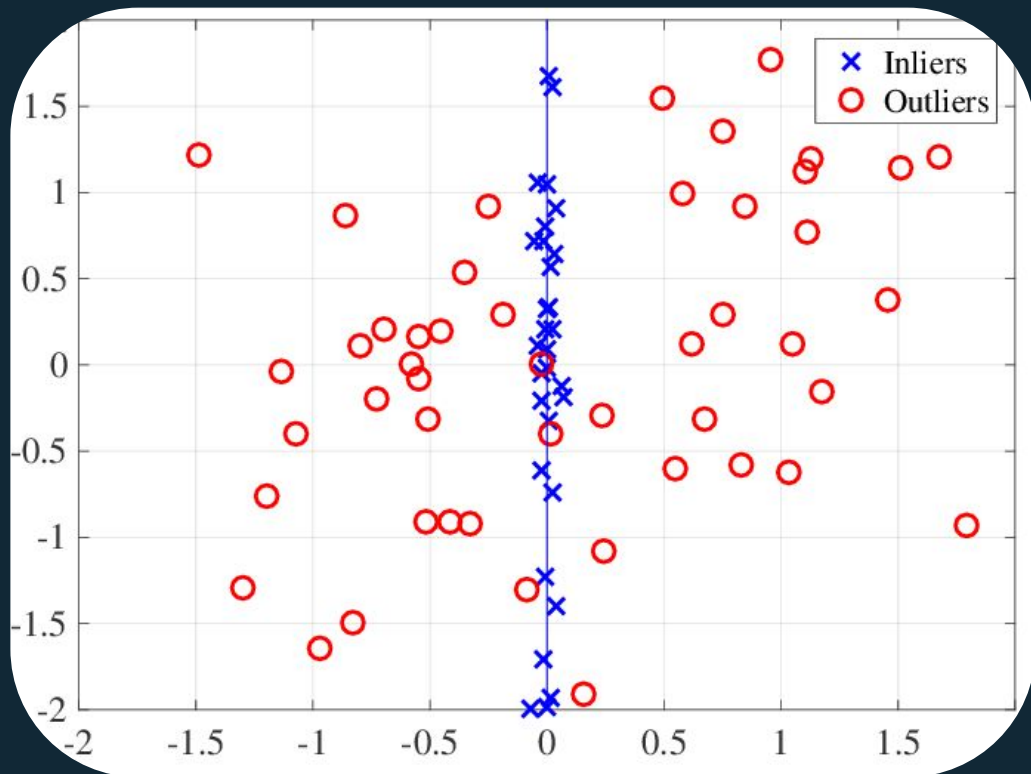
	Age	Gender	Salary	Department	Join_Date	Comments
0	35.0	M	60000.0	Finance	2018-01-07	Excellent
1	45.0	F	50000.0	Marketing	NaT	NaN
2	35.0	NaN	60000.0	NaN	2018-01-02	Good
3	28.0	M	60000.0	IT	2018-01-06	Average
4	30.0	NaN	NaN	Marketing	2018-01-06	Excellent
5	35.0	F	120000.0	NaN	NaT	NaN



## 4. Outliers e inliers

Los outliers son datos que se alejan significativamente del resto del conjunto, ya sea por errores, rarezas o casos excepcionales.

Los inliers son datos que se ajustan bien a un modelo o que está dentro de un rango esperado según una relación o distribución establecida.



Edades de estudiantes:  
[ 20, 19, 22, 21, 20, 95 ]

Inlier

Outlier





# 5. Valores Prototípicos y por defecto

## 1 Valores Prototípicos

Los valores prototípicos se refieren a la identificación y selección de instancias o patrones que son representativos de la mayoría de los datos o de una clase específica.

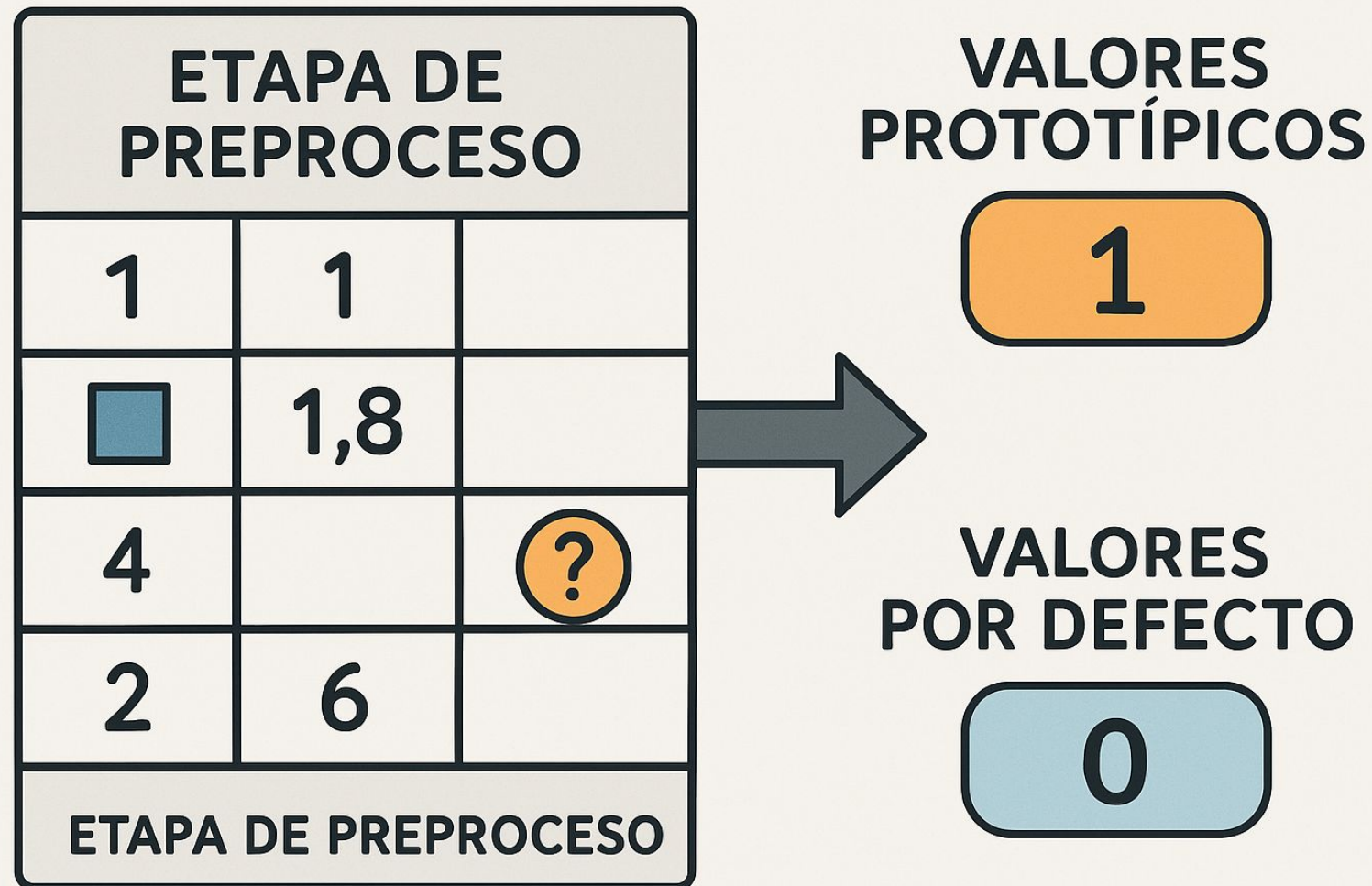
## 2 Valores por defecto

Por otro lado, los valores por defecto se utilizan para manejar los datos faltantes o ausentes en el conjunto de datos original.

ETAPA DE PREPROCESO		
1	1	
	1,8	
4		
2	6	
ETAPA DE PREPROCESO		



# Ejemplo de Valores Prototípicos y por defecto



Tenemos la siguiente tabla donde se están preprocesando los datos, vemos que hay valores faltantes o que no se han procesado correctamente. Para ello podemos usar o un valor prototípico como la media, mediana o moda, o un valor por defecto indicado anteriormente como 0, “Desconocido” o “NAN”.

## 6. Valores erróneos

Son datos considerados incorrectos, inconsistentes o fuera de lugar dentro de un conjunto de datos.

Estos pueden surgir por errores humanos, fallos a la hora de recolectarlos o problemas técnicos durante el proceso de captura o transmisión.



# Maneras de corregir los valores erróneos





Busca reducir la cantidad de datos, enfocándose en la información más relevante antes de aplicar los algoritmos de minería de datos.





# Beneficios de la agregación de datos

1

## Reducción de dimensionalidad

Se reduce la cantidad de información a procesar.

2

## Mejora de la calidad de los datos

Menor ruido o n° de valores atípicos.

3

## Facilidad en el análisis

Menos datos facilitan encontrar relaciones y patrones.

4

## Eficiencia computacional

Se opera más rápido y con menor consumo de recursos.



[illegible]

Puede llevar a la pérdida de información detallada. Por ello, debe hacerse de manera estratégica para no perder patrones relevantes que podrían ser útiles para el modelo o el análisis que se desea realizar.



# ¡Gracias por su Atención!

Esperamos que esta presentación haya sido útil.

