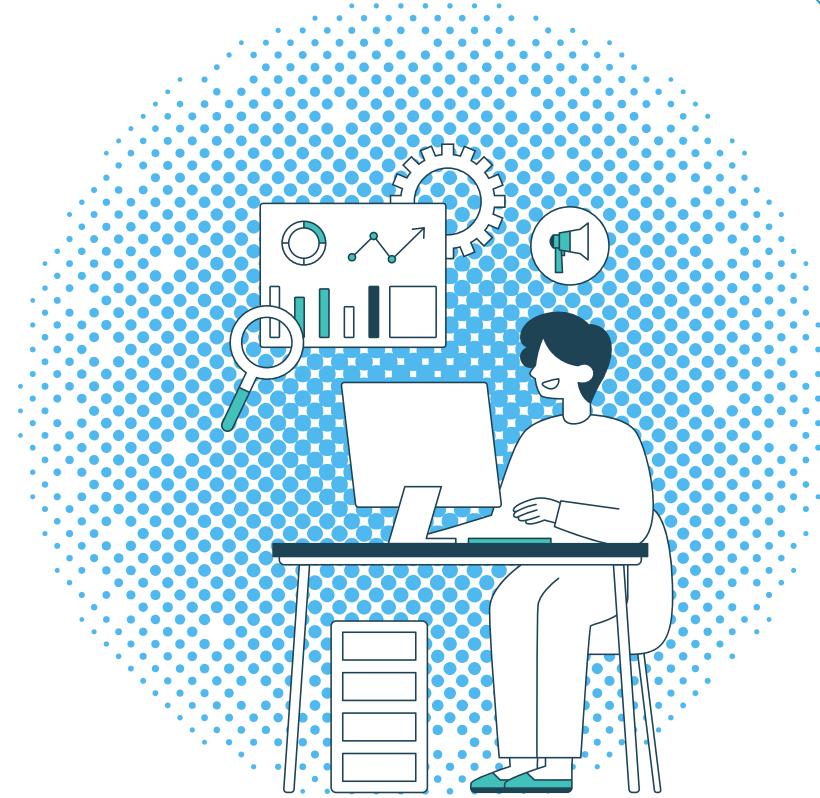


Análisis de datos

Carrera de Especialización en
Inteligencia Artificial



Programa de la materia

1 Introducción al análisis de datos

Flujo de trabajo típico. Aplicaciones. Configuración del entorno. Introducción a las principales bibliotecas.

2 Análisis exploratorio de datos

EDA, estadística descriptiva y visualización.

3 Preprocesamiento de datos y Feature Engineering

Tipos de variables. Codificación. Manejo de datos faltantes. Creación de nuevos features.

4 Pruebas estadísticas y validación de hipótesis

Pruebas de normalidad y de correlación. ANOVA y test de dependencia.

5 Reducción de dimensionalidad

Métodos para reducir la dimensionalidad y técnicas para selección de features.

6 Taller práctico

Análisis de datos completo de un dataset. Buenas prácticas. Discusión grupal.

7 Presentación de trabajos finales

Exposición por grupos y devolución de los docentes.

8 Automatización del análisis de datos

Herramientas para automatizar el EDA y otras tareas del flujo de trabajo.

Tipos de variables

Cosas que puedo contar o medir

Cuantitativas (numéricas)

Discretas

Toman valores enteros

Continuas

Pueden tomar cualquier valor dentro de un rango

Cantidad de hijos

Número de visitas al médico

Número de ventas

Temperatura

Altura

Edad

Cualidades o información que no se puede medir

Categóricas (cualitativas o atributos)

Nominales

Sin orden

Ordinales

Tienen un orden específico

Estado civil

Nombre de ciudad

Día de la semana

Tallas de ropa (S, M, L)

Opiniones (muy de acuerdo, de acuerdo,...)

Nivel de educación (secundario, terciario, universitario)

Algunas consideraciones interesantes

- La **edad** se considera una variable continua pero generalmente se mide y reporta como discreta (Ej., en años).
- Las variables **binarias** (Sí/No, 1/0, True/False) son un caso particular de las categóricas nominales.
- Las **fechas** tienen una doble naturaleza:
 - Numérica, cuando interesa medir tiempo transcurrido o tendencias.
 - Categórica, cuando interesa agrupar eventos en una fecha específica (Ej., analizar las ventas en un día particular de la semana).
- ¿Qué pasa con las **coordenadas geográficas**?



Exploración completa de datos



1. Exploración de datos numéricos (clase 2)



Tendencia central

1

Media: valor promedio.

2

Mediana: en un conjunto ordenado de números, es el valor que se encuentra en el medio.

3

Moda: valor más frecuente.

Dispersión

4

Varianza: dispersión de los datos con respecto a la media. Promedia diferencias al cuadrado.

5

Desviación estándar: dispersión con respecto a la media pero en las mismas unidades que los datos.

6

Cuartiles y rango intercuartil (IQR): otra forma de cuantificar dispersión de los datos.

Forma

7

Curtosis y skewness: describe la forma de la distribución (simetría y la presencia de colas más pronunciadas o aplanadas).

2. Exploración de datos categóricos

- Se centra en conteos y proporciones.
- Medidas Principales:
 - Frecuencias absolutas: número de veces que aparece cada categoría.
 - Frecuencias relativas: porcentaje o proporción de cada una.
 - Moda: la categoría más frecuente.
- Visualizaciones:
 - Gráficos de barras: muestran la frecuencia de cada categoría.
 - Gráficos de torta: explican las proporciones relativas.
 - Tablas de contingencia (cross-tab): revelan la relación entre dos variables categóricas.



Ejemplos prácticos en Jupyter



Preparación y transformación de los datos

La mayoría de los algoritmos de ML no pueden procesar los datos a menos que estén “limpios” .

- Manejo de datos faltantes.
- Variables numéricas: outliers, escala (normalización/estandarización).
- Variables categóricas: balance, cardinalidad y codificación.
- Relaciones entre variables.



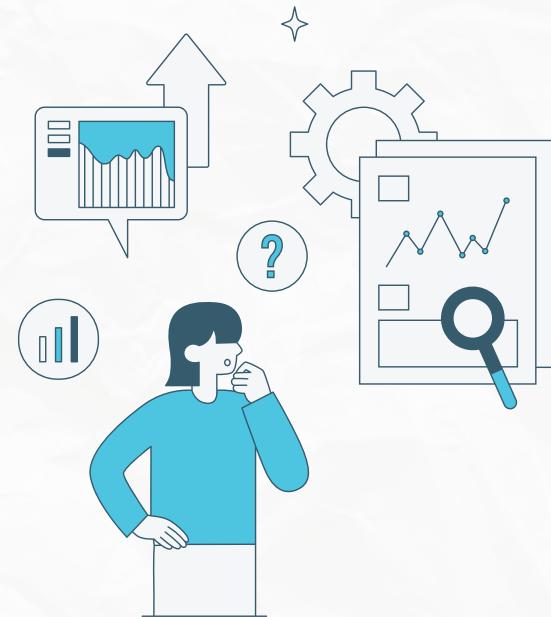
¿Qué son los datos faltantes?



- Son observaciones incompletas o valores no registrados en el dataset (NaN, null).
- Pueden distorsionar el análisis y los modelos si no se tratan (introducir bias y afectar la confiabilidad).
- Existen distintas estrategias para atacar el problema.
- Entender el motivo de ausencia de los datos es importante, porque permite seleccionar la estrategia adecuada.

Tipos de causas de datos faltantes

- **MCAR (Missing Completely at Random)**: el valor falta por una causa al azar y su falta es independiente de los datos.
- **MAR (Missing at Random)**: la falta del valor se relaciona con los datos de otras columnas.
- **MNAR (Missing Not at Random)**: la falta depende de los datos no observados y se debe a una causa específica que debe ser investigada.



Ejemplo 1

Ejemplo: estamos investigando salarios en Argentina y enviamos una encuesta a 100 personas.

CASO 1: las personas se olvidaron de responder preguntas al azar.

Participante	Edad	Rubro	Salario
1	29	Abogado/a	1.500.000
2	53	Null	430.000
3	21	Analista de datos	Null
4	34	Analista de marketing	1.000.000
5	Null	A. financiero/a	600.000
5	31	Asesor/a del congreso	700.000
6	48	Recepcionista	Null
.....

La falta no depende de los datos observados, no hay ningún patrón.

MCAR?

Ejemplo 2

Ejemplo: estamos investigando salarios en Argentina y enviamos una encuesta a 100 personas.

CASO 2: los entrevistados mayores de 40 no se sienten cómodos de divulgar su salario.

Participante	Edad	Rubro	Salario
1	29	Abogado/a	1.500.000
2	37	Profesor/a	430.000
3	53	Analista de datos	Null
4	34	Analista de marketing	1.000.000
5	23	A. financiero/a	600.000
5	31	Asesor/a del congreso	700.000
6	48	Recepcionista	Null
.....

Hay un patrón, la falta depende de los datos observados (la edad).

MAR?

Ejemplo 3

Ejemplo: estamos investigando salarios en Argentina y enviamos una encuesta a 100 personas.

CASO 3: las personas con salarios más altos tienden a evitar contestar la pregunta.

Participante	Edad	Rubro	Salario
1	45	Abogado/a	Null
2	52	Profesor/a	430.000
3	21	Analista de datos	Null
4	34	Analista de marketing	Null
5	23	A. financiero/a	600.000
5	31	Asesor/a del congreso	700.000
6	48	Recepcionista	770.000
.....

La falta depende del dato que no tengo

MNAR?

Algunas formas de analizar causas

- **Métodos gráficos** para analizar si hay o no patrones.
- **Regresión logística** (ver si se predice ausencia de datos a partir de otras variables).
- **Test de Little** (basado en Chi cuadrado):
 - H₀, son MCAR
 - H₁, no son MCAR.
 - Desventajas:
 - Rechazar H₀ no indica si es MAR o MNAR.
 - No rechazar H₀ tampoco confirma MCAR.
 - El test asume distribución Normal.
 - No aplica para datos categóricos.

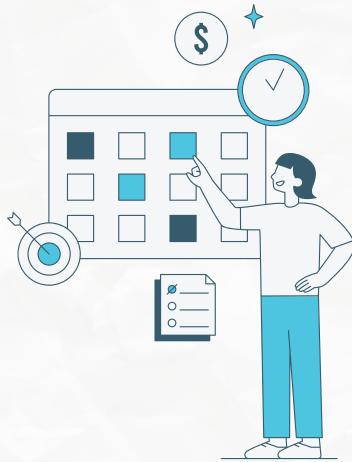


En definitiva...



- No podemos saber a ciencia cierta qué mecanismo gobierna la falta de datos.
- Con un buen análisis y entendimiento del dominio se pueden hacer hipótesis razonables.

Estrategias para manejo de datos faltantes



Eliminación:

- Borrar filas.
- Borrar columnas (variables).

Imputación (colocar un valor):

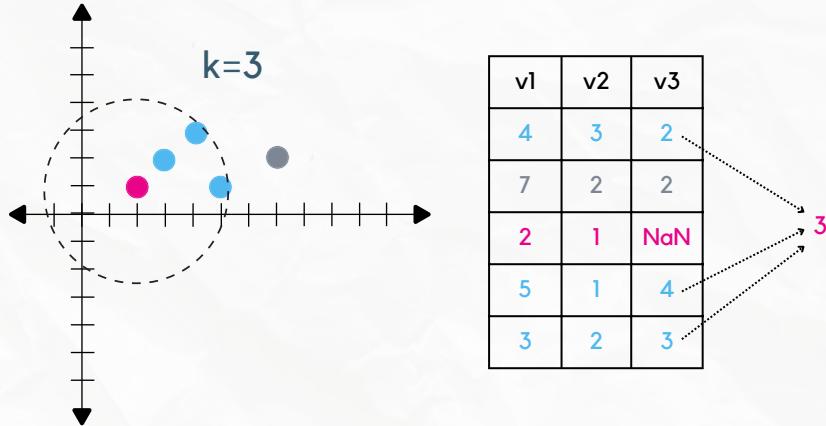
- Métodos univariados (se usan los datos de la variable con faltantes).
 - Por media, mediana, moda.
 - Valor de adelante o de atrás.
 - Crear nueva categoría.
 - Series temporales: hacia adelante/atrás e interpolación lineal o polinómica.
- Métodos multivariados (se usan otras columnas): usar modelos estadísticos (ej., **MICE**) y algoritmos de ML (por ej., regresión, árboles de decisión, KNN).

Métodos multivariados - KNN

KNN (K-Nearest Neighbors):
busca los "k" vecinos que se
parecen más a la fila del dato
faltante. Este se calcula luego a
partir de esos vecinos.

Atención:

- Se recomienda normalizar los datos
antes de usar este algoritmo porque es
sensible a la escala.
- Puede ser ineficiente en grandes
volúmenes de datos.



KNNImputer

Gallery examples: Release Highlights for scikit-learn 0.22 Imputing missing values before building an estimator

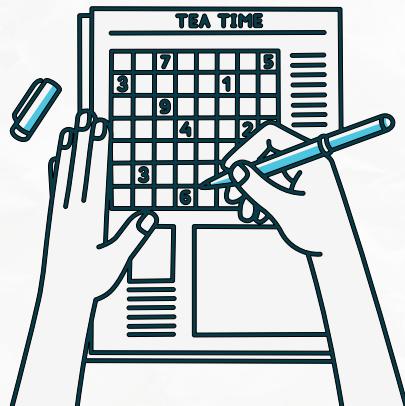


Métodos multivariados - MICE

MICE (Multiple Imputation by Chained Equations): es una técnica avanzada que predice valores faltantes por medio de modelar cada variable incompleta usando las otras columnas. Crea muchas versiones del dataset con distintas imputaciones que se pueden usar para un análisis estadístico más robusto.

Funcionamiento:

- Comienza con una imputación rápida y simple (Ej., la media).
- Iteración: Para cada columna con NaNs, predice los valores en base a las otras variables.
- Hace varias iteraciones, hasta que los valores predichos quedan estables.



Mice imputation (statsmodels)

IterativeImputer

Gallery examples: Imputing missing values before building an estimator
Imputing missing values with variants of IterativeImputer



(Parecido a MICE, no es igual)

¿Cuándo aplicar cada una? 1/2

- Si los faltantes son MCAR:
 - Eliminación de fila cuando los datos faltantes representan 5-10% máx. del total de filas.
 - Eliminación de columna: >60-70% valores faltantes en la misma columna y la variable no es importante.
 - Atención: evaluar el impacto en la distribución de los datos antes de eliminar filas. Eliminar datos puede producir sesgo.
- Si son MCAR o MAR y el porcentaje de faltantes es bajo (10-30%):
 - Imputación por media: datos con distribución aprox. Normal.
 - Imputación por mediana o KNN (captura mejor relaciones no lineales pero depende de la calidad de otras variables y puede implicar más costo computacional).
 - Imputación por moda (variables categóricas) o nueva categoría (ej., "FALTANTE") si tiene sentido.
 - Atención: imputar con la media, mediana y moda modifican la varianza.

¿Cuándo aplicar cada una? 2/2

- Si son MAR con patrones claros:
 - Imputación con modelos de ML para variables críticas.
 - Considerar algoritmos como Random Forest Imputer o KNN-imputer si los patrones son no lineales.
 - Porcentaje de faltantes moderado a alto (20-50%).
 - Atención: si la variable es muy crítica y los faltantes superan el 40-50%, se requiere un análisis más cuidadoso para evitar overfitting.
- Sospecha de MNAR o MAR complejos (dependen de muchas variables):
 - Multiple Imputation by Chained Equations (MICE).
- Series temporales: se elige una imputación específica según el tipo de serie (interpolación, modelos de series temporales como ARIMA, LSTMs, modelos generativos).