CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

Diapositivas 03 · Imputación, pipelines y efecto sobre modelos

Tratamientos reproducibles, prevención de leakage y evaluación del impacto en modelos y decisiones.



Módulo 02 · Calidad de datos y valores faltantes

Material elaborado por el profesor Sergio Gevatschnaider

IDEA CLAVE

Imputar es modelar

Cada reemplazo introduce una hipótesis sobre el dato que no se observó.



La imputación no recupera la verdad: construye una estimación útil bajo supuestos.

Por eso debe evaluarse por su efecto en distribución, relaciones, modelo y decisión.

COMPARACIÓN

Estrategias de tratamiento

Ninguna técnica es universal: todas cambian algo.

Estrategia	Ventaja	Riesgo
Eliminar filas	Simple y transparente.	Pierde casos y puede sesgar.
Media global	Rápida para variables numéricas.	Reduce varianza y oculta segmentos.
Mediana	Robusta a extremos.	Aplana la distribución.
Media por grupo	Conserva diferencias observables.	Depende de grupos confiables.
Indicador de ausencia	Captura señal informativa.	Puede sobreajustar si hay leakage.
Modelo/pipeline	Reproducible y escalable.	Debe validarse estrictamente.

MARCO DE TRABAJO

Dónde ocurre el error más frecuente

Imputar antes de dividir datos contamina la evaluación.

Flujo incorrecto

- Calcular imputador con toda la base.
- Separar train/test después.
- Evaluar sobre datos indirectamente vistos.
- Reportar una métrica optimista.

Flujo correcto

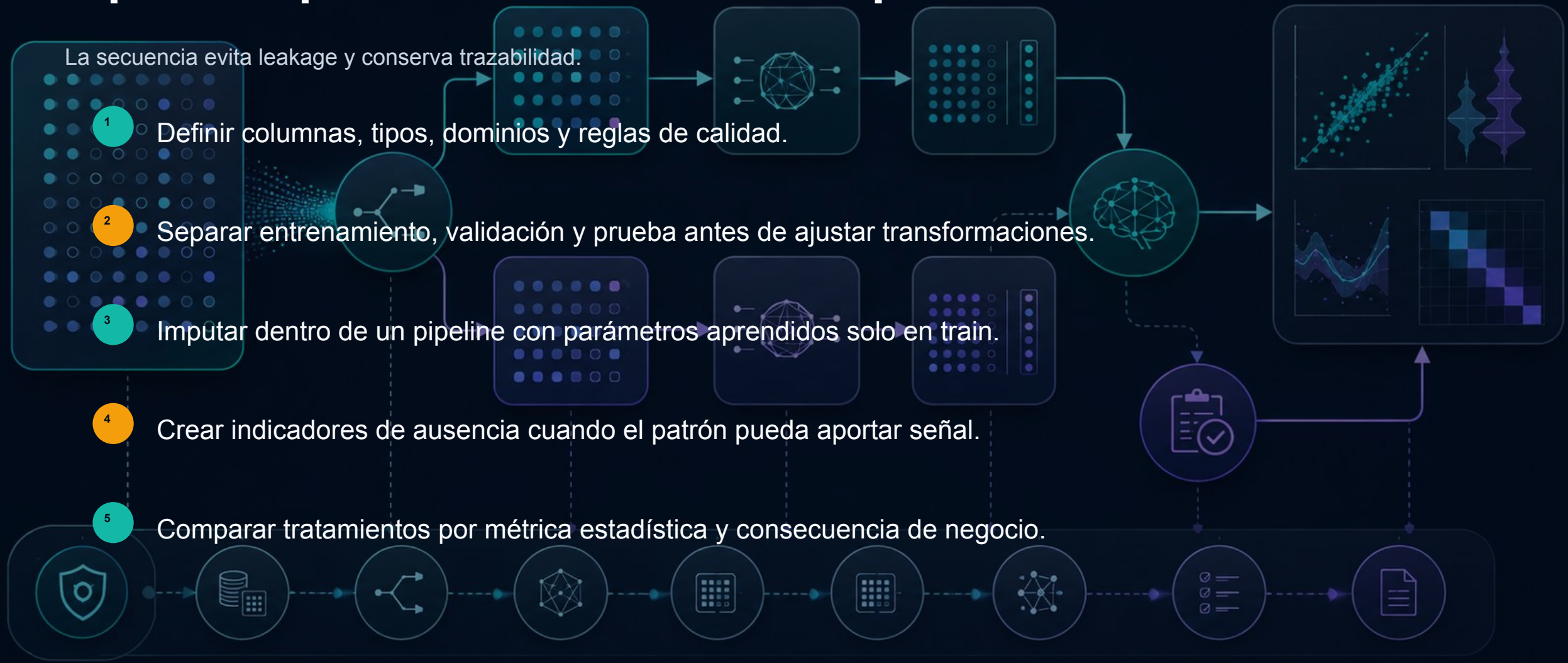
- Separar train/test primero.
- Ajustar imputador solo en train.
- Aplicar la misma transformación a test.
- Guardar pipeline reproducible.

SECUENCIA

Pipeline reproducible de calidad e imputación

La secuencia evita leakage y conserva trazabilidad.

1. Definir columnas, tipos, dominios y reglas de calidad.
2. Separar entrenamiento, validación y prueba antes de ajustar transformaciones.
3. Imputar dentro de un pipeline con parámetros aprendidos solo en train.
4. Crear indicadores de ausencia cuando el patrón pueda aportar señal.
5. Comparar tratamientos por métrica estadística y consecuencia de negocio.



CONCEPTO

Caso House Prices: el faltante puede tener significado

No toda celda vacía indica error de captura.

- Variables como garaje, pileta, chimenea o sótano pueden faltar porque el atributo no existe.
- Imputar con promedio puede borrar una señal estructural relevante para precio.
- Codificar “no aplica” o crear indicadores puede mejorar interpretación y desempeño.
- El tratamiento correcto nace del significado del dato, no de una receta automática.



MARCO DE TRABAJO

Indicadores de ausencia

A veces el hecho de que falte un dato informa más que el valor imputado.

Cuándo ayudan

- Faltante asociado a proceso, segmento o preferencia.
- Atributo inexistente con impacto en la decisión.
- MNAR probable o señal operacional.
- Modelos capaces de usar interacciones.

Cuándo dañan

- Ausencia puramente aleatoria.
- Muestra pequeña con muchos indicadores.
- Patrón cambia entre train y producción.
- La ausencia revela información no disponible al momento de decidir.

COMPARACIÓN

Evaluar más allá del accuracy

La imputación puede mejorar una métrica y empeorar la decisión.

Evidencia	Qué mirar	Pregunta
Distribución	Media, mediana, varianza, extremos.	¿Se deformó la población?
Relaciones	Correlaciones e interacciones.	¿Cambió la estructura?
Modelo	RMSE, MAE, AUC, calibración.	¿Generaliza o solo ajusta?
Segmento	Errores por grupo.	¿Perjudica a un subgrupo?
Decisión	Costo, riesgo, umbral.	¿Cambia la acción?
Operación	Monitoreo y drift.	¿Se mantiene en producción?

CONCEPTO

Cómo documentar la recomendación

Una decisión defendible separa hechos, supuestos y límites.

- Hecho observado: porcentaje y patrón de faltantes por variable y segmento.
- Hipótesis: mecanismo probable y justificación.
- Tratamientos comparados: eliminación, imputación simple, grupo, indicador o modelo.
- Impacto: sesgo, dispersión, métrica final y costo de decisión.
- Límite: qué no puede afirmarse y qué dato conviene recolectar.



IDEA CLAVE

Principio de cierre

La técnica debe estar subordinada a la decisión.

El mejor tratamiento es el que conserva evidencia suficiente para actuar con riesgo explícito.

El producto final del módulo debe permitir que otra persona replique el diagnóstico, entienda el supuesto y cuestione la recomendación.

