CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

Diapositivas 01 · Calidad de datos y decisiones

Dimensiones de calidad, gobernanza, trazabilidad y costo de error antes de modelar.

Módulo 02 · Calidad de datos y valores faltantes

Material elaborado por el profesor Sergio Gevatschnaider

IDEA CLAVE

La calidad no es un requisito técnico

Es una condición para que la evidencia represente el proceso real.

Una decisión solo es tan sólida como la evidencia que la sostiene.

El módulo parte de una pregunta profesional: qué defectos de datos pueden cambiar la conclusión y cuáles pueden documentarse, corregirse o bloquear la decisión.



SECUENCIA

De dato registrado a decisión defendible

El recorrido analítico debe conservar trazabilidad de cada transformación.

1. Proceso real: cliente, operación, sensor, formulario o sistema transaccional.
2. Captura: reglas, campos obligatorios, formatos, integraciones y tiempos.
3. Diagnóstico: completitud, validez, consistencia, unicidad y oportunidad.
4. Tratamiento: imputar, eliminar, corregir, segmentar o recolectar nuevamente.
5. Decisión: comunicar impacto, límite, costo esperado y responsable.

COMPARACIÓN

Dimensiones de calidad de datos

Completitud es solo una de las dimensiones que pueden modificar una decisión.

Dimensión	Pregunta	Riesgo
Completitud	¿Qué información falta y en qué segmento?	Muestra no representativa.
Validez	¿El valor respeta formato, rango y dominio?	Reglas de negocio violadas.
Consistencia	¿Coincide entre fuentes y momentos?	Métricas contradictorias.
Unicidad	¿Hay duplicados reales o registros repetidos?	Conteos y tasas infladas.
Oportunidad	¿El dato llega a tiempo para decidir?	Acción tardía o incorrecta.
Trazabilidad	¿Se puede auditar la regla aplicada?	Resultado no defendible.

MARCO DE TRABAJO

Defecto tolerable vs. defecto bloqueante

La severidad depende de la decisión, no del porcentaje aislado de error.

Admite corrección

- Baja incidencia y patrón explicable.
- Regla de negocio documentada.
- Impacto pequeño en la métrica final.
- El tratamiento se replica en train/test.

Bloquea la decisión

- Sesgo concentrado en un grupo clave.
- Variables críticas ausentes o no confiables.
- La conclusión cambia entre tratamientos.
- No existe forma de auditar el origen.

CONCEPTO

Calidad como contrato entre negocio y datos

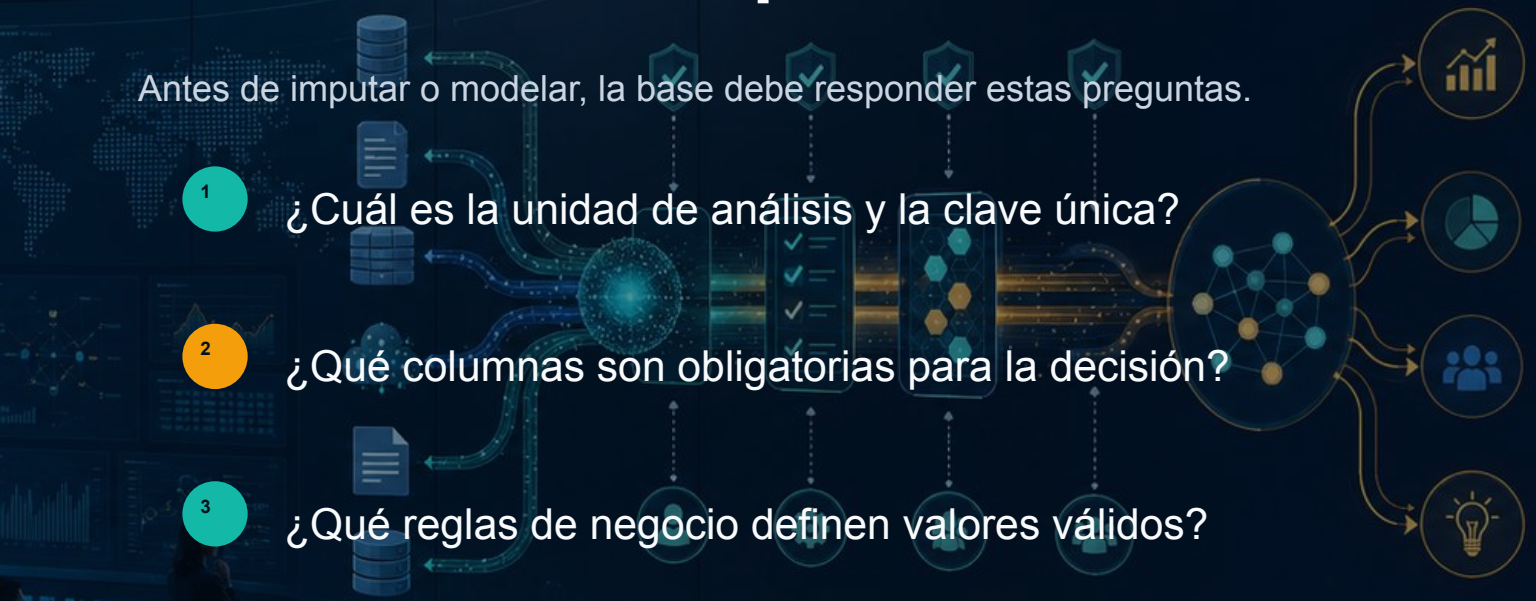
Las reglas deben expresarse antes del análisis, no después del resultado.

- Cada variable crítica necesita definición, unidad, rango válido y fuente autorizada.
- Cada regla debe tener severidad: advertencia, corrección automática, revisión o bloqueo.
- Cada excepción debe conservarse: dato original, regla aplicada, fecha, responsable y justificación.
- La calidad operativa se monitorea como cualquier proceso productivo.



SECUENCIA

Checklist mínimo para auditar una base



Antes de imputar o modelar, la base debe responder estas preguntas.

1. ¿Cuál es la unidad de análisis y la clave única?
2. ¿Qué columnas son obligatorias para la decisión?
3. ¿Qué reglas de negocio definen valores válidos?
4. ¿Dónde se concentran faltantes, duplicados o inconsistencias?
5. ¿Qué decisión cambia si el tratamiento cambia?

MARCO DE TRABAJO

Costo estadístico y costo de negocio

Un error pequeño puede ser grave si cruza un umbral operativo.

Costo estadístico

- Sesgo de media, tasa o proporción.
- Pérdida de varianza real.
- Correlaciones alteradas.
- Métrica predictiva sobreestimada.

Costo de negocio

- Clientes mal segmentados.
- Crédito aprobado o rechazado injustamente.
- Inventario asignado al canal equivocado.
- Forecast y presupuesto fuera de rango.

CONCEPTO

Gobernanza: quién decide sobre la calidad

La limpieza sin responsable convierte el análisis en una caja negra.

- Owner del dato: define significado, origen y regla de negocio.
- Analista: mide impacto, compara tratamientos y documenta supuestos.
- Área usuaria: decide si el riesgo residual es aceptable para actuar.
- Auditoría/model risk: exige trazabilidad, reproducibilidad y límites explícitos.



IDEA CLAVE

Principio de cierre

La recomendación siempre debe incluir un límite.

No alcanza con “limpiar datos”. Hay que demostrar que la decisión sigue siendo válida.

El entregable del módulo no es una base corregida, sino un reporte de calidad con reglas, evidencia, impacto y próximo dato necesario.