Análisis Predictivo y Gestión de Datos

Sesión 2: Preparación y Limpieza de Datos

Oscar Leonardo Rincón León

24 de abril de 2025

Objetivos de la sesión

- Comprender cómo la calidad de los datos afecta la capacidad predictiva
- Identificar problemas comunes en los datos y su impacto en los modelos
- Conectar la limpieza con la etapa de preparación de datos en CRISP-DM
- Aplicar estrategias prácticas de depuración y transformación

¿Por qué importa la preparación en análisis predictivo?

- Los modelos aprenden patrones: si los datos están distorsionados, aprenderán mal
- Variables mal estructuradas $\rightarrow$ patrones falsos
- Ruido $=$ reducción del poder predictivo
- La limpieza es una forma de modelar: decidir qué mantener y cómo

¿Qué es CRISP-DM?

- **CRISP-DM** significa Cross-Industry Standard Process for Data Mining.
- Es un modelo estándar y ampliamente adoptado para desarrollar proyectos de ciencia de datos.
- Proporciona una estructura organizada para abordar problemas de negocio utilizando datos.

¿Por qué usar CRISP-DM?

- Es aplicable en múltiples sectores: salud, educación, industria, gobierno.
- Su enfoque cíclico permite iterar, corregir y mejorar continuamente.
- Ayuda a comunicar el proceso de análisis entre equipos multidisciplinarios.
- Aporta una guía paso a paso para estructurar el trabajo analítico desde la comprensión del problema hasta la implementación.

Etapas del ciclo CRISP-DM (1/2)

- **1. Comprensión del contexto:** entender el problema, sus objetivos y su impacto en un área como educación, salud o administración pública.
- **2. Comprensión de los datos:** recolectar, explorar y validar la calidad de los datos disponibles.
- **3. Preparación de los datos:** limpiar, transformar, codificar y seleccionar variables útiles para el análisis posterior.

Etapas del ciclo CRISP-DM (2/2)

- **4. Modelado:** seleccionar y entrenar modelos estadísticos o de aprendizaje automático.
- **5. Evaluación:** verificar si el modelo resuelve el problema, usando métricas adecuadas y revisando supuestos.
- **6. Despliegue:** comunicar resultados, implementar soluciones o integrarlas en procesos de decisión.

Preparación en el ciclo CRISP-DM

- La preparación de datos es la **tercera etapa** del ciclo
- Asegura que los datos estén limpios, completos y listos para el modelado
- Impacta directamente en:
  - Selección adecuada de modelos
  - Métricas válidas de evaluación
  - Interpretación clara de resultados

Decisiones clave al preparar datos

- ¿Qué hacer con valores faltantes?
- ¿Qué variables eliminar?
- ¿Qué transformar?
- ¿Qué codificar o escalar?
- Cada decisión afecta la precisión, estabilidad e interpretabilidad del modelo

Errores comunes y sus consecuencias

- Mantener variables con alta proporción de nulos $\rightarrow$ modelos inconsistentes
- No escalar variables $\rightarrow$ sesgo en modelos sensibles a magnitudes (KNN, SVM)
- Duplicados $\rightarrow$ sobreajuste
- Codificación incorrecta $\rightarrow$ distorsión de relaciones

¿Por qué es crucial la preparación de los datos?

- La calidad de los datos condiciona la calidad de los modelos predictivos.
- Los modelos aprenden patrones de los datos disponibles: si hay errores, aprenden mal.
- En proyectos reales, más del 70 % del tiempo se invierte en preparación, no modelado.

Preparación en el ciclo CRISP-DM

- Es la tercera etapa del ciclo y conecta la comprensión del negocio con la implementación del modelo.
- Transforma datos crudos en una base coherente y lista para el análisis.
- Afecta directamente la calidad de los resultados, su interpretación y utilidad práctica.

Errores comunes a corregir

- **Valores faltantes:** eliminación o imputación (media, mediana, por grupo).
- **Variables mal codificadas:** ordinales, categóricas, numéricas que no lo son.
- **Escalas incompatibles:** requiere escalado para evitar sesgos.
- **Outliers:** afectan modelos sensibles; deben detectarse y analizarse.

Preparación y aprendizaje supervisado

- En aprendizaje supervisado, todo modelo depende de una buena relación entre variables explicativas (X) y objetivo (Y).
- Si los datos no están bien preparados, los modelos no aprenderán patrones útiles.
- La limpieza garantiza interpretabilidad, generalización y confianza en los resultados.

Referencias clave

- **Hands-On ML (Cap. 2):** destaca la limpieza y transformación como parte esencial del flujo ML.
- **ISLP:** enfatiza que un análisis válido depende de datos estructurados y bien documentados.
- **CRISP-DM:** define la preparación como puente entre exploración y modelado.

Preparación práctica con Python

- `pandas`: limpieza, filtrado, imputación
- `numpy`: operaciones matemáticas y máscaras
- `scikit-learn`: escalado, codificación, pipelines

Tratamiento de valores faltantes

- Opciones:
  - Eliminación: si son pocos y aleatorios
  - Imputación: media, mediana, moda, por grupos
  - Modelos de imputación (avanzado)
- Toda imputación introduce supuestos

Codificación de variables categóricas

- Convertir texto en números para modelos
- Opciones:
  - Ordinal: cuando hay jerarquía
  - One-hot encoding: para evitar supuestos
- Evitar codificación que induzca relaciones falsas

Escalado de variables numéricas

- Modelos como KNN y regresión logística son sensibles a magnitudes
- Opciones:
  - Min-Max Scaling
  - Z-score (estandarización)
- Escalar antes de entrenar y validar, nunca después

Ejercicio práctico

- Dataset con problemas reales:
  - Valores nulos
  - Categóricas no codificadas
  - Magnitudes no comparables
- Meta: preparar para su uso en un modelo de clasificación binaria

Cierre de la sesión

- La preparación no es técnica: es estratégica
- Afecta todo el proceso de modelado
- Un analista predictivo prepara pensando en el modelo, el error y la interpretación

Tarea: Aplicar limpieza a un nuevo dataset. Documentar decisiones y justificar imputaciones.