Análisis Predictivo y Gestión de Datos

Sesión 3: Exploración y Visualización de Patrones

Oscar Leonardo Rincón León

24 de abril de 2025

Objetivo de la sesión

Analizar visualmente la estructura de los datos para guiar la selección de modelos predictivos.

¿Por qué es importante la exploración de datos (EDA)?

- Permite comprender la estructura y comportamiento de los datos.
- Detecta errores, valores atípicos y relaciones inesperadas.
- Guía la selección del modelo predictivo y el preprocesamiento.
- Fortalece la interpretación de los resultados.

Elementos clave en EDA

- **Distribuciones univariadas:** histogramas, boxplots, KDE.
- **Relaciones bivariadas:** scatterplots, boxplots por categoría.
- **Correlaciones:** mapas de calor.
- **Outliers y valores nulos:** detección visual y estadística.
- **Transformaciones:** log, normalización, escalado.

Distribuciones y visualización univariada

- Ayudan a identificar asimetrías, colas largas, sesgo.
- La elección de bins en histogramas influye en la percepción.
- Boxplots muestran mediana, cuartiles y valores atípicos.
- KDE suaviza la distribución (estimación de densidad).

Relaciones bivariadas y correlaciones

- **Diagramas de dispersión:** relaciones entre dos variables numéricas.
- **Mapas de calor:** matriz de correlación visual.
- **Boxplots por grupo:** para comparar variables numéricas entre categorías.
- Correlación $\neq$ causalidad, pero puede guiar hipótesis.

Outliers y estructura de los datos

- Valores extremos pueden ser errores, casos reales o mal codificados.
- Su detección es clave para evitar distorsión en modelos.
- La visualización facilita su identificación antes del análisis.
- Algunos algoritmos son sensibles a outliers (ej. KNN, regresión).

Visualización efectiva: más que estética

- El objetivo no es embellecer, sino informar y guiar decisiones.
- Cada gráfico debe responder a una pregunta específica.
- Buenas prácticas: claridad, etiquetas, escala apropiada.
- La visualización también comunica hallazgos al público no técnico.

Herramientas prácticas en Python

- `matplotlib`: base para gráficos personalizados.
- `seaborn`: enfoque estadístico y estético para gráficos rápidos.
- `pandas`: soporte básico para gráficos integrados en DataFrames.

Actividad práctica

- Usar el dataset limpio con 1300 registros.
- Visualizar:
  - Histogramas y boxplots univariados.
  - Diagramas de dispersión y mapas de calor.
- Detectar patrones visuales útiles para la predicción.
- Reflexionar sobre posibles transformaciones.

Cierre de la sesión

- El EDA no es opcional: mejora la comprensión y orienta decisiones.
- Una visualización bien hecha permite anticipar errores, confirmar hipótesis o replantear modelos.
- El análisis visual es tanto exploratorio como comunicativo.