Análisis exploratorio de datos

M4: Análisis exploratorio de datos

|AE1: Aplicar análisis exploratorio de datos distinguiendo sus principales técnicas y herramientas.

Introducción👩‍💻

En el mundo actual, donde los datos son uno de los activos más importantes para la toma de decisiones, resulta fundamental conocer y dominar técnicas que permitan explorarlos, comprenderlos y validarlos antes de cualquier análisis complejo o modelado.

El **Análisis Exploratorio de Datos (EDA)** representa una etapa crítica en el flujo de trabajo de la ciencia de datos, ya que permite identificar patrones, detectar errores, plantear hipótesis y establecer relaciones entre variables, todo a partir de una mirada preliminar e intuitiva sobre la información disponible.

Este manual tiene como objetivo introducir los conceptos fundamentales del EDA, diferenciándolo del Análisis Inicial de Datos (IDA), y presentar las principales **técnicas univariadas y multivariadas**, así como herramientas modernas utilizadas en entornos reales. A lo largo del documento, también se analizarán casos prácticos que facilitarán la aplicación concreta de los conocimientos adquiridos.

Aprendizaje esperado

Cuando finalices la lección serás capaz de:

- Comprender el propósito del análisis exploratorio de datos (EDA) y su valor en el proceso analítico.
- Distinguir claramente entre análisis inicial de datos (IDA) y análisis exploratorio.
- Identificar y aplicar técnicas univariadas y multivariadas para obtener insights a partir de datos reales.
- Reconocer herramientas y recursos útiles para ejecutar análisis exploratorios con eficiencia y claridad.
- Interpretar visualizaciones y estadísticas descriptivas para fundamentar decisiones basadas en datos.

¿Qué es el Análisis Exploratorio de Datos (EDA)?

El Análisis Exploratorio de Datos (EDA) es una metodología estadística que busca describir las características principales de un conjunto de datos utilizando métodos visuales y cuantitativos. Su propósito es obtener una comprensión inicial de los datos, sin suposiciones previas, con el fin de identificar patrones, tendencias, relaciones, valores atípicos y errores. A diferencia de los enfoques confirmatorios, el EDA es inductivo y abierto, lo que permite que el analista explore libremente el comportamiento de las variables antes de tomar decisiones sobre modelado o inferencia.

El EDA suele ser la primera etapa en proyectos de análisis, ciencia de datos o aprendizaje automático. En lugar de aplicar modelos directamente, se examinan gráficos, resúmenes numéricos y relaciones entre variables para comprender la estructura de los datos. Entre sus herramientas comunes se encuentran los histogramas, diagramas de caja (boxplots), gráficos de dispersión, medidas de tendencia central y matrices de correlación. Estas técnicas permiten formular hipótesis preliminares, verificar la validez de los datos y orientar las siguientes etapas del análisis.

En la práctica, el EDA es fundamental para evitar errores en fases posteriores. Por ejemplo, al observar un gráfico de dispersión se puede detectar una relación no lineal que requerirá transformar una variable. O al aplicar un histograma se puede evidenciar una distribución sesgada que justifica el uso de métodos robustos. El valor del EDA no está solo en mostrar lo evidente, sino en **descubrir lo inesperado**, como diría John Tukey, su principal impulsor. Su rol es clave para garantizar que los modelos y conclusiones estén bien fundamentados en la realidad de los datos.

¿Qué es el Análisis Inicial de Datos (IDA)?

El Análisis Inicial de Datos (IDA) es la etapa previa al análisis exploratorio, cuyo objetivo principal es **garantizar que los datos sean aptos para su análisis posterior**. Esta fase se enfoca en aspectos técnicos como la estructura del

dataset, la calidad de los registros, los tipos de datos y la presencia de valores faltantes o atípicos extremos. A diferencia del EDA, que busca interpretar los datos, el IDA se concentra en su validación, limpieza y transformación inicial para asegurar que los resultados futuros no se vean sesgados por errores u omisiones.

Entre las tareas clave del IDA se incluyen: la verificación del formato correcto de cada columna (numérica, categórica, fecha, etc.), la detección y cuantificación de valores nulos, la identificación de inconsistencias como duplicados o errores de codificación, y la detección de outliers extremos que puedan alterar la interpretación. Asimismo, es común generar reportes automáticos con herramientas como pandas-profiling o sweetviz que permiten inspeccionar rápidamente las principales características estructurales del conjunto de datos.

Un ejemplo práctico de IDA puede verse en un proyecto de análisis de clientes, donde el campo "edad" aparece con registros como "150", o el campo "email" contiene valores nulos o mal formateados. Si no se detectan y corrigen estos errores, cualquier visualización o modelo posterior estaría sesgado o directamente equivocado. Por eso, el IDA **actúa como una etapa de control de calidad**, necesaria para comenzar el análisis con datos confiables y correctamente estructurados.

Contexto en el cual se utiliza el Análisis Exploratorio de Datos

El Análisis Exploratorio de Datos (EDA) se aplica en múltiples etapas de un proyecto de datos y en una variedad de industrias. Desde investigaciones científicas hasta campañas de marketing digital, el EDA permite **comprender la información antes de tomar decisiones importantes**. Es una herramienta útil para generar hipótesis, identificar tendencias preliminares y orientar el análisis estadístico o predictivo. Su flexibilidad lo hace indispensable tanto en contextos académicos como empresariales.

En la práctica profesional, el EDA suele utilizarse al inicio de proyectos de ciencia de datos, machine learning o business intelligence. Por ejemplo, en un entorno bancario, se aplica EDA para revisar el comportamiento histórico de los clientes y detectar patrones de incumplimiento de pagos. En medicina, se usa para explorar variables clínicas antes de diseñar modelos diagnósticos. También en recursos humanos puede emplearse para visualizar correlaciones entre rotación de personal y condiciones laborales.

Más allá del ámbito, el EDA **funciona como puente entre los datos en crudo y la interpretación analítica**. Su rol no es confirmar hipótesis sino descubrir relaciones inesperadas, ayudando al analista a entender la lógica del fenómeno subyacente. Además, permite adaptar el tipo de visualizaciones y técnicas estadísticas que se utilizarán después, ajustando el enfoque según lo observado.

👉 *Ejemplo práctico: al explorar un dataset de consumo eléctrico, se puede observar una caída sistemática los fines de semana, lo que sugiere una relación temporal útil para análisis posteriores.*

Técnicas y herramientas para el análisis exploratorio

Las técnicas del EDA pueden dividirse en dos grandes grupos: estadísticas descriptivas y visualización de datos. Las primeras permiten resumir la información mediante medidas como media, mediana, moda, desviación estándar, percentiles, valores máximos y mínimos. Estas medidas ayudan a describir la forma y dispersión de los datos, y se aplican especialmente en el análisis univariado. Las visualizaciones, por otro lado, aportan una comprensión intuitiva y permiten detectar patrones no visibles en tablas de números.

Entre las visualizaciones más utilizadas están los histogramas (para conocer la distribución de variables numéricas), los diagramas de caja o boxplots (para observar la dispersión y outliers), los gráficos de barras (para comparar frecuencias de categorías), y los gráficos de dispersión (para explorar relaciones entre dos variables). También se emplean mapas de calor, diagramas de violín, gráficos de líneas temporales y matrices de correlación. Estas técnicas se combinan para ofrecer un panorama integral de los datos.

Las herramientas más comunes para realizar EDA son:

- **Python** con librerías como pandas, matplotlib, seaborn, plotly, sweetviz, pandas-profiling.
- **R y RStudio**, con funciones como summary(), ggplot2, dplyr y tidyr.
- **Excel**, Power BI o Tableau, para exploración visual empresarial.

La elección de herramientas depende del perfil del analista, la complejidad del proyecto y el volumen de datos. Lo importante es seleccionar técnicas adecuadas al tipo de variables y objetivos del análisis.

Análisis univariado y sus objetivos

El análisis univariado consiste en examinar una única variable de forma independiente, sin relacionarla con otras. Es el primer paso del EDA y su objetivo es **comprender la distribución, tipo y comportamiento general** de cada variable. Ayuda a identificar si una variable es numérica, categórica, ordinal o binaria, y qué tipo de medidas y gráficos pueden aplicarse para su estudio.

En el caso de variables numéricas, se analizan medidas como media, mediana, rango, desviación estándar y percentiles. También se utilizan histogramas para observar la forma de la distribución (simétrica, sesgada, con outliers) y boxplots para detectar valores extremos. Para variables categóricas, se revisan frecuencias absolutas y relativas, y se visualizan con gráficos de barras o pie charts. Estas técnicas permiten entender cuán representativa o balanceada es una categoría en el conjunto.

Aplicar un análisis univariado es clave para evitar errores posteriores. Por ejemplo, si una variable presenta muchos valores nulos o un único valor dominante, podría no ser útil en un modelo predictivo. Asimismo, detectar asimetrías en la distribución puede justificar la transformación de la variable (logaritmo, raíz, etc.). Este análisis simple es esencial para decidir qué variables conservar, transformar o eliminar, y permite ganar una comprensión precisa del dataset.

Cierre ✍️

El análisis exploratorio de datos (EDA) es mucho más que una etapa preliminar: es una herramienta esencial para desarrollar una comprensión profunda, crítica y visual de cualquier conjunto de datos. Lejos de limitarse a estadísticas simples, el EDA permite detectar errores, descubrir relaciones significativas, plantear hipótesis y tomar decisiones informadas antes de realizar modelados o reportes formales. Esta exploración inicial sienta las bases para un análisis más robusto y confiable.

Distinguir correctamente entre el **análisis inicial de datos (IDA)** y el **EDA** permite estructurar mejor el proceso de trabajo: primero asegurar la calidad y limpieza del dataset, y luego aplicar técnicas exploratorias para obtener valor e interpretar su estructura interna. Este orden lógico y metodológico permite evitar errores frecuentes cómo usar datos incompletos, interpretar mal una variable o construir modelos basados en suposiciones falsas.

Finalmente, dominar tanto el análisis **univariado** como el **multivariado**, y utilizar herramientas como Python, R o Power BI de forma adecuada, permite a los profesionales del dato desenvolverse con solvencia en proyectos reales. El EDA no es solo una etapa técnica: es un espacio para pensar, visualizar y conectar con los datos. Aplicarlo correctamente es una muestra de madurez profesional en el análisis y gestión de la información.

¡Nos vemos de nuevo, más adelante! 👋

Referencias📚

- Seaborn Developers. (2024). *Seaborn documentation*. https://seaborn.pydata.org
- Pandas Development Team. (2024). *Pandas documentation*. https://pandas.pydata.org
- Wickham, H. (2014). *Tidy data*. *Journal of Statistical Software*, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10
- Nuclio Digital School. (2023, marzo 9). *Exploratory Data Analysis (EDA): Qué es y cómo aplicarlo en ciencia de datos*. *Nuclio School Blog*. https://nuclio.school/blog/eda-exploratory-data-analysis/

¡Muchas gracias!

Nos vemos en la próxima lección