Análisis Exploratorio de Datos (EDA)

El “reconocimiento del terreno” antes de entrar en combate con modelos.

Gabriel Rengifo

Proceso proyectos ML / Data

MODELO/ DASHBOARD
Implementación y ajuste del modelo/Dashboard

ANALISIS EXPLORATORIO:
Limpieza, preparación, transformación.
Selección de variables (atributos).

DATOS:
Materia prima.

PREGUNTA SMART

¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *

RETO DE LA ORGANIZACIÓN:

DESPLIEGUE:
- Portal, empresa
- Modificación proceso
- Recurso humano

EVALUACION:
¿Resuelve la pregunta?
Validación criterios de éxito.

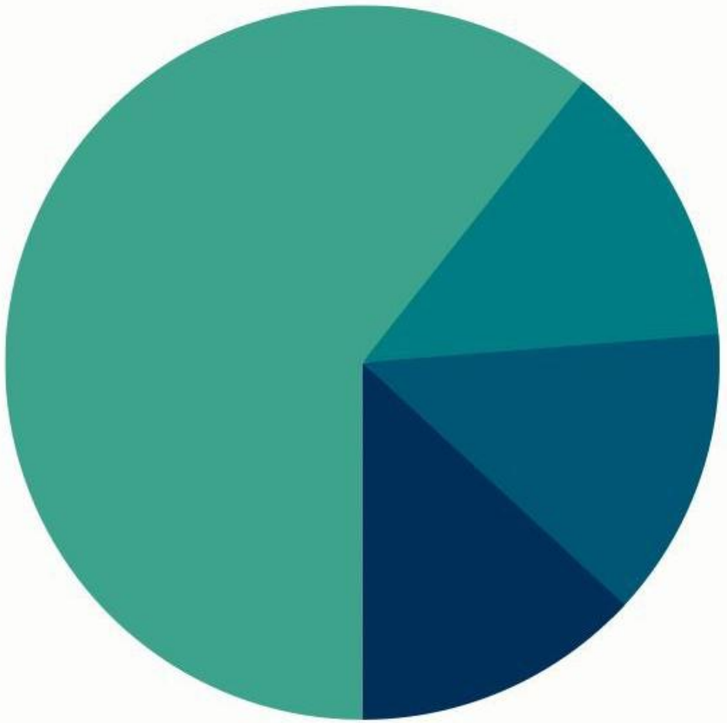
Data Analysts spend 60% of their time cleaning data

60%
DATA CLEANING

14%
DATA GATHERING

13%
DATA ANALYSIS

13%
OTHER TASKS

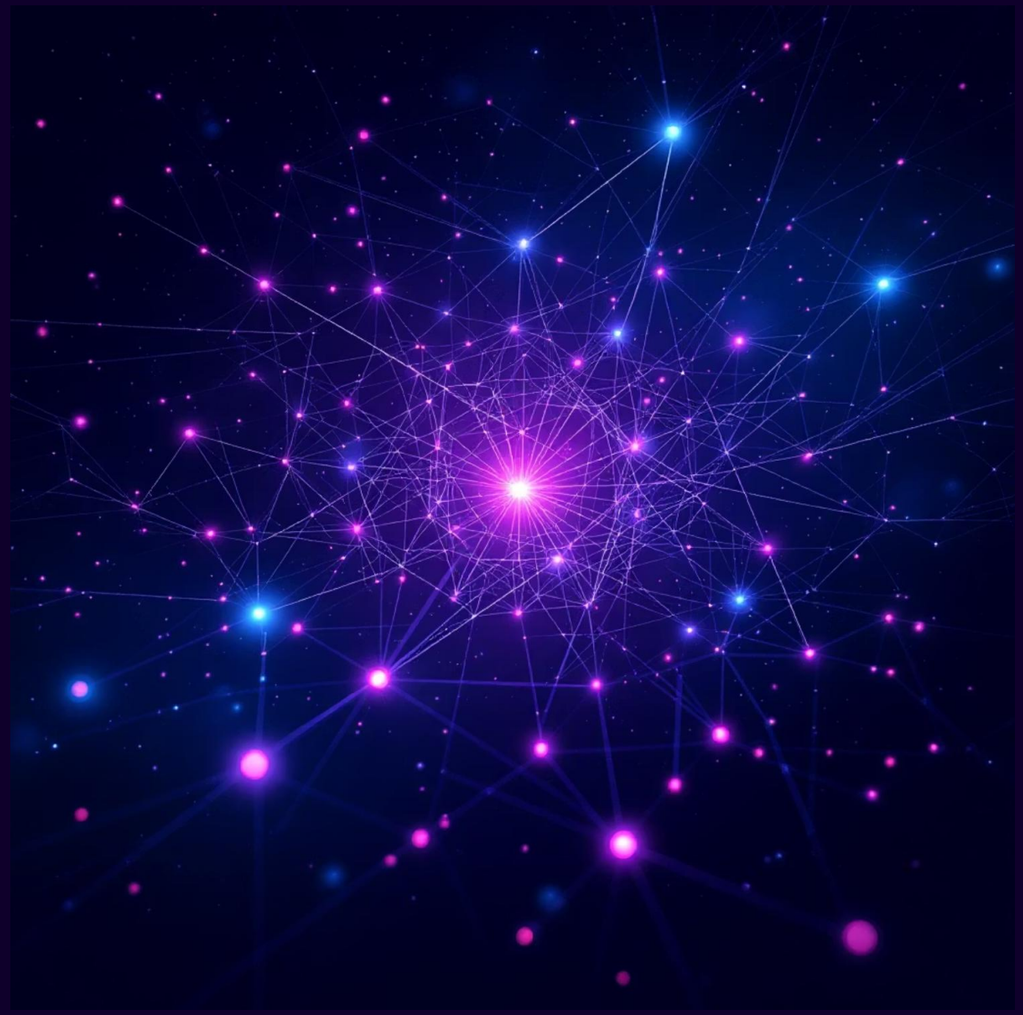


¿Qué es el Análisis Exploratorio de Datos (EDA)?

El EDA es el proceso inicial y fundamental de análisis de un conjunto de datos. Su objetivo principal es comprender la estructura, calidad y patrones intrínsecos de los datos.

Preguntas clave durante el EDA:

- ¿Qué variables componen el conjunto de datos?
- ¿Cuál es la completitud de la información disponible?
- ¿Existen anomalías, inconsistencias o valores atípicos?
- ¿Qué relaciones y correlaciones significativas se observan entre las variables?



Sin un EDA riguroso, cualquier modelo predictivo puede llevar a conclusiones y decisiones erróneas.

La Importancia Crucial del EDA en el Ámbito de la Defensa

Decisiones Críticas

Las operaciones militares y estratégicas exigen datos confiables y precisos para mitigar riesgos.

Ejemplos de Aplicación

- **Patrullajes:** Identificación predictiva de zonas de alto riesgo.
- **Sensores Navales:** Detección temprana de anomalías en buques para prevenir fallas críticas.
- **Comunicaciones:** Descubrimiento de patrones en flujos de datos para anticipar alertas.

Primera Línea de Defensa

El EDA actúa como el primer baluarte contra errores en los datos, asegurando la integridad de la información antes de su uso en modelos complejos.

El Proceso del EDA: Un Flujo de Trabajo Sistemático

01

1. Carga y Examen Inicial

Importar y obtener una primera visión del conjunto de datos.

02

2. Revisión Estructural

Analizar dimensiones (filas, columnas) y tipos de variables.

03

3. Identificación de Ausencias

Detectar y cuantificar valores nulos o registros duplicados.

04

4. Análisis Univariado

Estudiar cada variable de forma individual (distribuciones, estadísticas).

05

5. Análisis Multivariado

Explorar relaciones y correlaciones entre dos o más variables.

06

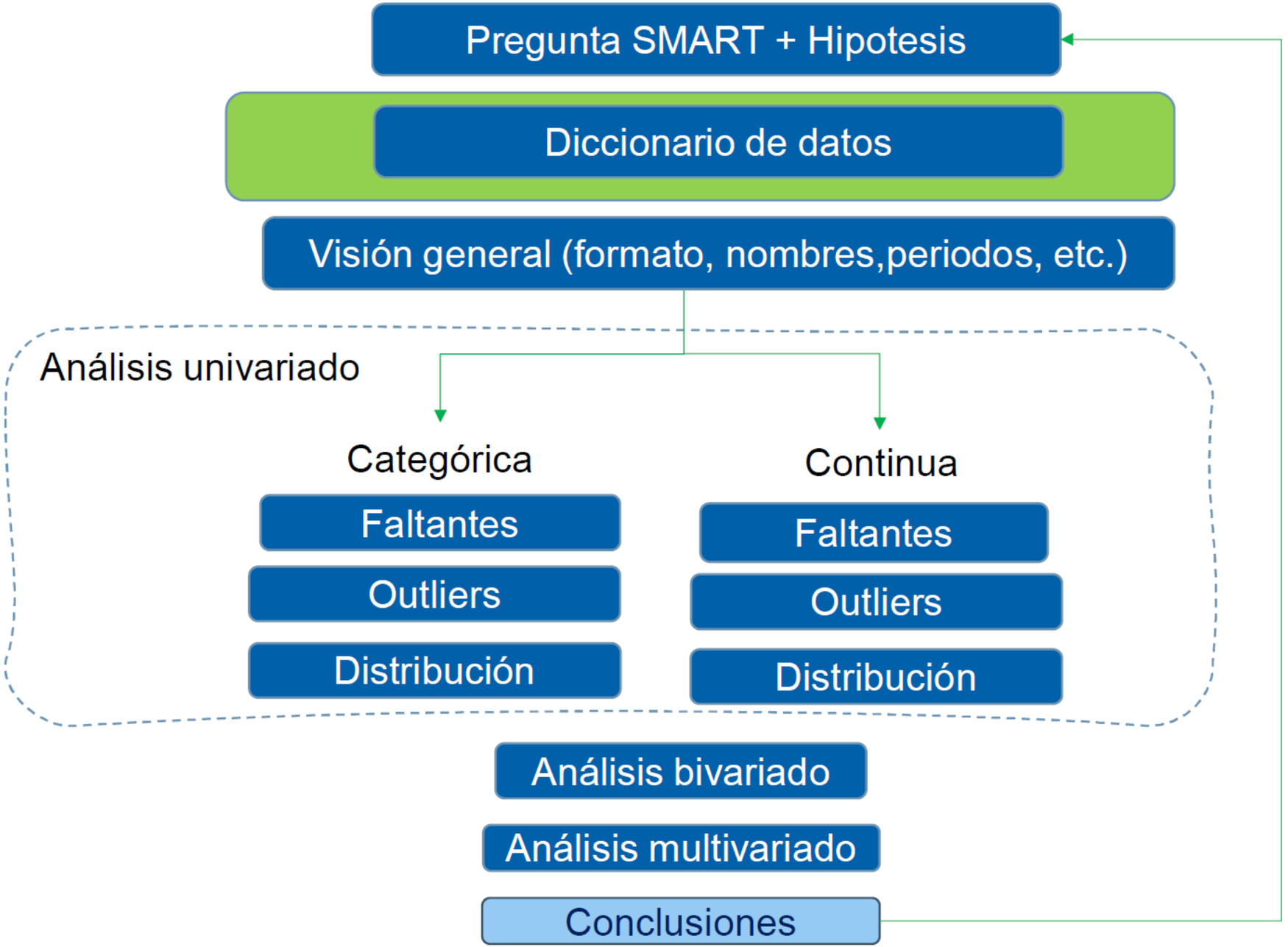
6. Detección de Anomalías

Identificar valores atípicos (outliers) y posibles errores en los datos.

07

7. Resumen y Preparación

Extraer hallazgos clave y documentar las necesidades de limpieza.


Pregunta SMART + Hipotesis
Diccionario de datos
Visión general (formato, nombres,periodos, etc.)
Análisis univariado
Categórica
Faltantes
Outliers
Distribución
Continua
Faltantes
Outliers
Distribución
Análisis bivariado
Análisis multivariado
Conclusiones

Diccionario de datos

Estructura base:

- **Variables:** Nombre de la variable.
- **Tipo:** Tipo o formato de la variable. Esto puede ser categórico, numérico, booleano, etc.
- **Contexto**: Información útil para entender el espacio semántico de la variable. En el caso de nuestro conjunto de datos, el contexto siempre es químico-físico, por lo que es fácil. En otro contexto, por ejemplo el inmobiliario, una variable podría pertenecer a un segmento en particular, como la anatomía del material o el social (¿cuántos vecinos hay?)
- **Expectativa**: ¿Qué tan relevante es esta variable con respecto a nuestra tarea? Podemos usar una escala “Alto, Medio, Bajo”.

Visión general de los datos

1. Formato de los datos
2. Nombre de las columnas (Nomenclatura)
3. Tipos de datos
4. Unidades de medida
5. Variables redundantes o no útiles (IDs e información personal, data leakage)
6. Nuevas columnas necesarias

Dataset de Práctica: El Titanic

Utilizaremos el clásico dataset del Titanic, una fuente de datos rica y compleja ideal para practicar el EDA.

- **Diversidad de Variables:** Contiene variables numéricas, categóricas y de texto.
- **Desafíos del Mundo Real:** Incluye valores faltantes y diversas anomalías que simulan problemas de datos reales, convirtiéndolo en un escenario perfecto para ejercicios de limpieza.

> **Pregunta de Investigación Central:**
>
> ¿Qué factores socioeconómicos y demográficos influyeron significativamente en la supervivencia de los pasajeros del Titanic?

Analogía Naval: La supervivencia en el Titanic puede verse como el éxito de una misión crítica, donde la comprensión de los factores es clave.

Exploración Inicial del Dataset

El primer contacto con los datos es crucial. Utilizamos las funciones básicas de la librería `pandas` para una visión general.

- **Herramientas Clave:** `df.head()` para las primeras filas, `df.info()` para la estructura y tipos de datos, y `df.describe()` para estadísticas descriptivas.
- **Objetivo Inmediato:** Determinar las **dimensiones** del dataset, verificar los **tipos de datos** inferidos para cada columna y observar las **primeras filas** para entender el formato de los valores.
- **Detección de Nulos:** Es vital revisar la distribución de valores nulos desde el inicio para identificar columnas incompletas.

```
# Ejemplo prácticoimport pandas as pddf = pd.read_csv('titanic.csv')print(df.info())print(df.isnull().sum())
```

Basado en esta primera exploración, ¿qué variables parecen requerir mayor atención o "limpieza" posterior?

Análisis Univariado: Comprendiendo Cada Variable

El análisis univariado se centra en el estudio de una única variable a la vez para entender su distribución y características individuales.

Variables Categóricas

- **Conteo de Valores:** Utilizar `value_counts()` para obtener la frecuencia de cada categoría.
- **Visualización:** Los gráficos de barras son ideales para representar estas frecuencias, facilitando la identificación de categorías dominantes o raras.

Variables Numéricas

- **Distribución:** Los **histogramas** muestran la forma de la distribución de los datos (simétrica, sesgada, multimodal).
- **Outliers y Dispersión:** Los **boxplots** son excelentes para identificar valores atípicos y la dispersión de los datos alrededor de la mediana.

Ejemplo práctico: Al analizar la distribución de edades en el dataset del Titanic, se pueden detectar valores extremos o la presencia de datos inusuales (e.g., edades superiores a 70 años).

Análisis Bivariado: Relaciones entre Variables

El análisis bivariado explora las **relaciones entre dos variables**, permitiendo descubrir patrones y dependencias que no son evidentes en el análisis univariado.

Sexo vs. Supervivencia

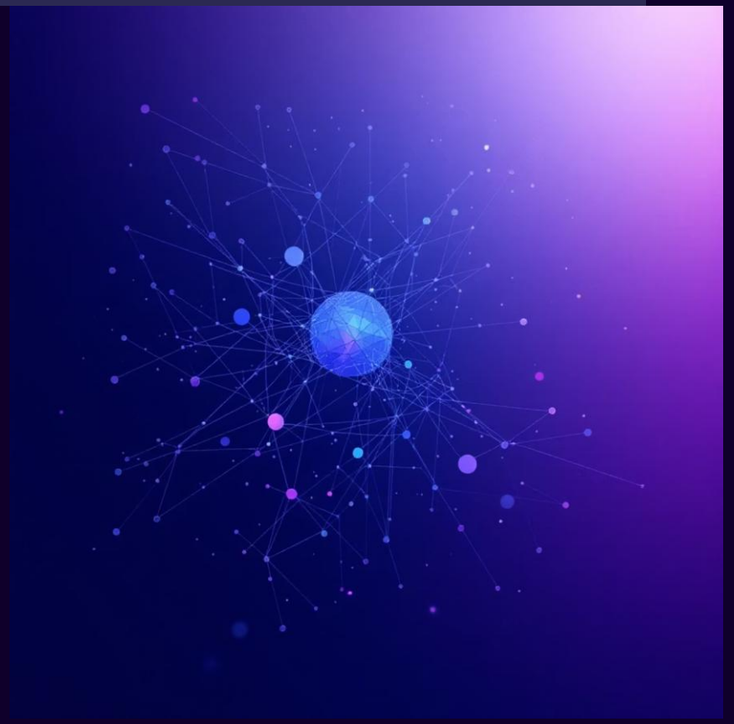
Análisis de la influencia del género en la tasa de supervivencia. Por ejemplo, en el Titanic, las mujeres mostraron una probabilidad significativamente mayor de sobrevivir.

Clase de Pasajero vs. Supervivencia

Estudio de cómo la clase socioeconómica (1ra, 2da, 3ra) se correlaciona con las tasas de supervivencia.

Técnicas Comunes:

- **Tablas de Contingencia (`crosstab`):** Cruce de dos variables categóricas para ver frecuencias conjuntas.
- **Gráficos de Dispersión:** Para visualizar la relación entre dos variables numéricas.
- **Correlaciones:** Cuantificación de la fuerza y dirección de la relación lineal entre variables numéricas.



Este análisis es fundamental para **priorizar** las variables que probablemente serán más predictoras en la construcción de modelos.

Hallazgos Clave del EDA en el Dataset del Titanic

Tras un proceso de EDA exhaustivo, hemos identificado varios puntos críticos en el dataset del Titanic que requieren atención antes del modelado:

1. Datos Faltantes Significativos

Variables como `Cabin` y `Age` presentan un alto porcentaje de valores nulos, lo que podría requerir estrategias de imputación o eliminación.

2. Disparidad en Supervivencia por Género

Las mujeres y los niños tuvieron una tasa de supervivencia significativamente mayor, lo que sugiere un factor cultural de "mujeres y niños primero".

3. Influencia de la Clase Social

Los pasajeros de clase alta (1ra clase) mostraron una mejor tasa de supervivencia, probablemente debido al acceso a botes salvavidas.

4. Outliers en la Tarifa

La variable `Fare` (tarifa) presenta valores atípicos, con algunos pasajes con costos extremadamente altos que podrían distorsionar análisis futuros.

Conclusión Crucial: Antes de cualquier intento de modelado predictivo, es imprescindible llevar a cabo una fase de limpieza y normalización profunda de los datos.

Resumen y Próximos Pasos

Lo Aprendido Hoy

- Definición y criticidad del **Análisis Exploratorio de Datos** (EDA).
- Cómo realizar una **exploración inicial** de un dataset en bruto.
- Aplicación de **técnicas univariadas y bivariadas** para el análisis.
- Importancia de la **identificación de valores nulos, atípicos** y descubrimiento de patrones.

Tarea para la Próxima Sesión

- Crear un **notebook** (Jupyter/Colab) que muestre al menos **3 hallazgos significativos** adicionales del dataset del Titanic mediante análisis univariado o bivariado.
- Subir el notebook a su **repositorio personal de Git** (GitHub/GitLab).

Próxima Clase: Profundizaremos en las técnicas de **limpieza y preprocesamiento de datos**, abordando los desafíos identificados hoy.