
Material de apoyo
Semillero de investigación:
Inteligencia Artificial de las Cosas
(AIoT)



Taller EDA

Tecnología en Desarrollo de Software -
Ingeniería de Sistemas y
Telecomunicaciones
Departamento de Ciencias Básicas e
Ingeniería

Fecha límite: —

28 de abril de 2025

Taller práctico

1. Introducción Teórica

1.1. ¿Qué es EDA?

El **Análisis Exploratorio de Datos (EDA)** es el proceso inicial de examinar los datos para descubrir patrones, anomalías, relaciones y probar hipótesis utilizando estadísticas y visualizaciones. Su objetivo es **entender** los datos antes de aplicar modelos predictivos.

1.2. Aplicaciones reales

- Comprensión de bases de datos en investigación científica.
- Preparación de datos en proyectos de Machine Learning.
- Análisis de métricas de negocio para estrategias de marketing.

1.3. Errores comunes

- No manejar valores nulos correctamente.
- Sacar conclusiones precipitadas sin análisis profundo.
- Ignorar la escala de los datos al comparar variables.

2. Herramientas de EDA en Python

- **Pandas:** manipulación de datos tabulares.
- **Matplotlib y Seaborn:** visualización de datos.
- **Numpy:** operaciones matemáticas eficientes.

3. Ejemplo práctico paso a paso

Supongamos que analizamos un conjunto de datos de "Ventas de productos".

```
# Paso 1: Importar librerías
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Paso 2: Cargar datos
df = pd.read_csv('ventas.csv')

# Paso 3: Exploración inicial
print(df.head())
print(df.info())
print(df.describe())

# Paso 4: Detección de valores nulos
print(df.isnull().sum())

# Paso 5: Estadísticas básicas
print(df['precio'].mean())
print(df['producto'].value_counts())

# Paso 6: Visualización
sns.histplot(df['precio'], kde=True)
plt.title('Distribución de Precios')
plt.show()

sns.boxplot(x='categoria', y='precio', data=df)
plt.title('Precio por Categoría')
plt.show()
```

4. Ejercicio Propuesto

Conjunto de datos sugerido: **Iris dataset**.

1. Cargar el dataset.
2. Realizar un análisis de valores faltantes.
3. Crear histogramas para las variables numéricas.
4. Crear un gráfico de correlación (heatmap).
5. Identificar patrones o anomalías.

5. Visual Sugerido

Se recomienda incluir:

- Gráfico de histogramas por variable.
- Mapa de calor (heatmap) de correlaciones.