

Tarea 1: Análisis Exploratorio de Datos

AST332-1, Segundo Semestre 2025

Para esta tarea, usted deberá realizar un análisis exploratorio de datos (EDA) aplicando técnicas y métodos utilizados en clases. El objetivo es familiarizarse con el dataset, identificar patrones o anomalías, y obtener una comprensión general de los datos para futuras aplicaciones de machine learning.

Para realizar el análisis, debe escoger sólo un set de datos dentro de la siguiente lista:

- Películas y TV Shows de Disney Plus: <https://www.kaggle.com/datasets/unanimad/disney-plus-shows>
- Dataset Completo de Pokemon: <https://www.kaggle.com/datasets/rounakbanik/pokemon>
- Dataset Canciones de Spotify: <https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset>
- Dataset de Cambio Climático: <https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv>

Debe escoger sólo un set de datos!

Su análisis debe contener

0. Breve descripción del contexto de los datos:

Debe explicar los datos que escoge, la información de la columna, y cualquier antecedente importante para poder

1. Carga y descripción básica de los datos

- Cargar el dataset utilizando pandas y realizar una inspección inicial de las primeras filas.
- Mostrar información básica sobre las columnas (nombre, tipo de datos, valores nulos, etc.).

2. Limpieza de datos

- Identificar y manejar valores nulos o faltantes.
- Detectar y eliminar valores duplicados si los hubiera.
- Comentar sobre posibles outliers y cómo podrían influir en el análisis.

3. Análisis estadístico

- Calcular y mostrar estadísticas descriptivas de las variables numéricas (media, mediana, desviación estándar, etc.).
- Realizar un análisis de las distribuciones de las variables **más relevantes**.

4. Visualización de datos

- Crear gráficos para visualizar distribuciones, relaciones entre variables y patrones interesantes (histogramas, boxplots, scatter plots, etc.).
- Crear un heatmap para mostrar la correlación entre las variables numéricas.

5. Análisis de correlación

- Calcular la matriz de correlación y discutir sobre las relaciones encontradas entre las variables.
- Identificar cuáles variables parecen estar más relacionadas entre sí y qué implicaciones podría tener esto para futuros modelos de machine learning.

6. Conclusiones

- Resumir los hallazgos principales del análisis.
- Proponer hipótesis que podrían ser investigadas con modelos de machine learning en el futuro.

Instrucciones de entrega

- Usted puede trabajar solo o en grupo de máximo 2 integrantes.
- El trabajo debe ser entregado en un Jupyter Notebook (1 notebook por grupo), con el código y las visualizaciones claramente comentados.
- Además de los aspectos del análisis descritos más arriba, se evaluará la claridad de su notebook, que incluya comentarios que permitan una lectura fácil, y bibliografía usada, si aplica.
- Se habilitará un link para la entrega
- La fecha de entrega es el día **10 de Septiembre, hasta 23.59 hrs.**

Uso de IA generativa

Usted debe citar y dar crédito a cualquier interacción que tenga con Chatgpt o cualquier otra herramienta similar para la realización de esta tarea, mostrando en una sección "Apéndice" los prompts utilizados y las respuestas obtenidas. Debe explicar de qué forma utilizó la herramienta (generar ideas, generación de código, etc.) y la razón por la cual eligió utilizarla (ahorrar tiempo, mejorar un texto, bloqueo de ideas, etc.) **Recuerde que las herramientas IA generativas no siempre son confiables y usualmente cometen errores.** Usted debe ser crítico y evaluar las respuestas generadas. La responsabilidad del uso del contenido generado por estas herramientas es de exclusiva responsabilidad de la/el estudiante. La profesora se reserva el derecho de decidir acciones adicionales respecto al uso de AI.