ACTIVIDAD PRÁCTICA: ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Simulación de los 11 Pasos del EDA con Datasets Reales

I. INTRODUCCIÓN

Esta actividad tiene como objetivo que los estudiantes apliquen los 11 pasos fundamentales del Análisis Exploratorio de Datos (EDA) utilizando datasets reales. Cada grupo trabajará con un dataset diferente, siguiendo la misma metodología para garantizar la comparabilidad de resultados y el aprendizaje colaborativo.

II. OBJETIVOS DE APRENDIZAJE

- Aplicar técnicas de limpieza y preparación de datos
- Realizar análisis exploratorio mediante visualizaciones
- Identificar patrones, tendencias y anomalías en los datos
- Formular y responder preguntas de investigación basadas en datos
- Desarrollar modelos predictivos básicos
- Comunicar hallazgos de manera efectiva mediante presentaciones

III. LOS 11 PASOS DEL EDA A SEGUIR

Paso 1: Importar Librerías
Importar pandas, numpy, matplotlib, seaborn, sklearn y otras librerías necesarias.

Paso 2: Importar Datos
Cargar el dataset asignado y crear una copia de trabajo para preservar los datos originales.

Paso 3: Revisar Datos NAN
Identificar valores faltantes usando .info(), .describe(), .isna().sum() y visualizaciones con heatmap.

Paso 4: Limpiar Datos
Tratar valores faltantes, corregir inconsistencias, estandarizar categorías y eliminar duplicados.

Paso 5: Realizar Gráficas para Analizar Tendencias
Crear visualizaciones (countplot, histplot, boxplot, scatter, pie, etc.) para cada variable relevante y analizar cada gráfica.

Paso 6: Conversión de Variables a Número
Aplicar LabelEncoder o técnicas similares para convertir variables categóricas a numéricas.

Paso 7: Normalización
Normalizar los datos usando MinMaxScaler o StandardScaler para estandarizar las escalas.

Paso 8: Correlación
Calcular y visualizar la matriz de correlación (Pearson o Spearman) mediante heatmap.

Paso 9: Test de Normalidad
Aplicar el test de Shapiro-Wilk para verificar si las variables siguen distribución normal.

Paso 10: Pregunta de Investigación
Formular una pregunta de investigación específica basada en el análisis exploratorio realizado.

Paso 11: Ejercicio de Predicción
Implementar un modelo de regresión o clasificación para responder la pregunta de investigación, evaluar su rendimiento (MSE, R^2, accuracy, etc.) y visualizar resultados.

IV. DATASETS ASIGNADOS POR GRUPO

Grupo 1: Iris Dataset

- Descripción: Clasificación de especies de flores iris
- URL: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

Grupo 2: Wine Quality Dataset

- Descripción: Calidad de vinos tintos
- URL: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

Grupo 3: Diabetes Dataset

- Descripción: Predicción de diabetes en pacientes
- URL: https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv

Grupo 4: Boston Housing Dataset

- Descripción: Precios de viviendas en Boston
- URL: https://raw.githubusercontent.com/selva86/datasets/master/BostonHousing.csv

Grupo 5: Titanic Dataset

- Descripción: Supervivencia de pasajeros del Titanic
- URL: https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv

Grupo 6: Heart Disease Dataset

- Descripción: Predicción de enfermedades cardíacas
- URL: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

V. ENTREGABLES

A. Notebook de Python (.ipynb o .py)

Código completo con los 11 pasos del EDA, incluyendo comentarios explicativos en cada sección.

B. Presentación PowerPoint (8 slides)

La presentación debe contener exactamente 8 slides con la siguiente estructura:

Slide 1: Portada
Título del proyecto, nombre del dataset, integrantes del grupo, fecha

Slide 2: Pregunta de Investigación
Pregunta específica que se busca responder con el análisis del dataset

Slide 3: Análisis del Dataset
Descripción general: número de registros, variables, tipos de datos, estadísticos descriptivos

Slide 4: Visualizaciones Principales (Parte 1)
Incluir 2-3 gráficas clave con sus respectivos análisis

Slide 5: Visualizaciones Principales (Parte 2)
Incluir 2-3 gráficas adicionales con análisis de tendencias y patrones

Slide 6: Características Encontradas en el Proceso
Patrones identificados, correlaciones importantes, outliers, distribuciones relevantes

Slide 7: Predicción y Hallazgos
Resultados del modelo predictivo (métricas: R^2, MSE, accuracy), interpretación de coeficientes, respuesta a la pregunta de investigación

Slide 8: Limitaciones y Restricciones
Limitaciones del dataset, restricciones del análisis, posibles sesgos, recomendaciones para futuros análisis

C. Conclusiones Escritas

Documento de 1-2 páginas con conclusiones detalladas que respondan a la pregunta de investigación, incluyendo insights obtenidos, implicaciones prácticas y recomendaciones.

VI. CRITERIOS DE EVALUACIÓN

Criterio	Peso	Descripción
Implementación de los 11 Pasos del EDA	25%	Completitud y correcta aplicación de cada paso
Calidad de las Visualizaciones	15%	Gráficas claras, bien etiquetadas y con análisis pertinentes
Limpieza y Preparación de Datos	15%	Tratamiento adecuado de valores faltantes e inconsistencias
Pregunta de Investigación y Modelo Predictivo	20%	Pregunta bien formulada y modelo apropiado con evaluación correcta
Presentación PowerPoint	15%	Estructura clara, diseño profesional, contenido completo en 8 slides
Conclusiones y Hallazgos	10%	Análisis profundo, respuesta clara a la pregunta de investigación

VII. CRONOGRAMA

- Asignación de grupos y datasets: 1 de Diciembre de 2025
- Entrega final (Código + PPT + Conclusiones):
- Grupo 1 al 6 – 11 de Diciembre de 2025
- Presentaciones grupales:
- Grupo 1 al 6 – 12 de Diciembre de 2025

VIII. RECOMENDACIONES

- Trabajen en equipo y distribuyan las tareas equitativamente
- Documenten cada paso con comentarios claros en el código
- No se limiten a copiar código; entiendan cada línea
- Investiguen sobre el contexto de su dataset para hacer análisis más profundos
- Practiquen la presentación antes del día de exposición
- Consulten con el profesor ante dudas específicas
- Utilicen control de versiones (Git) para gestionar el código del equipo

IX. RECURSOS ADICIONALES

- Documentación de Pandas: https://pandas.pydata.org/docs/
- Documentación de Seaborn: https://seaborn.pydata.org/
- Documentación de Scikit-learn: https://scikit-learn.org/
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/
- Kaggle Learn: https://www.kaggle.com/learn