Actividad Práctica: Análisis e Interpretación de Datos con Python

Objetivo

Aplicar técnicas de análisis exploratorio, estadística descriptiva, distribuciones de probabilidad, clustering y reducción de dimensionalidad para extraer información útil de un conjunto de datos real.

Requisitos Previos

- Uso de Python y librerías: `pandas`, `numpy`, `matplotlib`, `seaborn`, `scipy.stats`, `sklearn`.
- Familiaridad con manipulación de datos, estadística básica y machine learning.
- Entorno recomendado: Anaconda y Jupyter Notebooks.

Conjunto de Datos

Usaremos el dataset `penguins` de la librería `seaborn`, que contiene características de tres especies de pingüinos.

Columnas principales:

- `species`: Especie del pingüino.
- `bill_length_mm`, `bill_depth_mm`: Longitud y profundidad del pico en mm.
- `flipper_length_mm`: Longitud de las aletas en mm.
- `body_mass_g`: Masa corporal en gramos.
- `island`: Isla de origen.
- `sex`: Sexo del pingüino.

Ejercicios

Ejercicio 1: Análisis Exploratorio y Estadísticos Básicos

Instrucciones:

1. Carga el conjunto de datos y visualiza las primeras filas.
2. Identifica los valores nulos y decide cómo imputarlos adecuadamente.
3. Calcula los estadísticos básicos (media, mediana, desviación estándar, percentiles) para las variables numéricas.
4. Visualiza la distribución de la masa corporal (`body_mass_g`) diferenciando entre especies usando un gráfico adecuado.

Ejercicio 2: Distribuciones de Probabilidad

Instrucciones:

1. Ajusta una distribución normal a la variable `body_mass_g` para cada especie.
2. Realiza un test de normalidad (como Shapiro-Wilk) para verificar si los datos siguen una distribución normal.
3. Grafica la distribución de los datos junto con la curva de la distribución normal ajustada.

Ejercicio 3: Clustering con K-Means

Instrucciones:

1. Selecciona las variables `bill_length_mm` y `bill_depth_mm` para realizar el clustering.
2. Aplica el algoritmo K-Means con 3 clusters.
3. Agrega la información de los clusters al DataFrame.
4. Visualiza los clusters obtenidos usando un gráfico de dispersión e identifica los centroides.

Ejercicio 4: Reducción de Dimensionalidad con PCA

Instrucciones:

1. Selecciona las variables numéricas del dataset y elimina las no numéricas.
2. Aplica PCA para reducir la dimensionalidad a 2 componentes principales.

3. Transforma los datos usando PCA y crea un nuevo DataFrame con las componentes principales.
4. Grafica los datos transformados, coloreando por especie.