MINERÍA DE DATOS Y MODELIZACIÓN PREDICTIVA

Actividad de ACP y Cluster

Descripción del Conjunto de Datos 'penguins'

El conjunto de datos `'penguins'` de la librería `'seaborn'` de Python contiene la siguiente información sobre diferentes especies de pingüinos:

- **species:** Es la especie de pingüino. Hay tres especies en el conjunto de datos: 'Adelie', 'Chinstrap' y 'Gentoo'.
- **island:** Representa la isla donde se recopilaron los datos. Las islas son 'Biscoe', 'Dream' y 'Torgersen'.
- **bill_length_mm:** Longitud del pico en milímetros.
- **bill_depth_mm:** Profundidad del pico en milímetros.
- **flipper_length_mm:** Longitud de la aleta en milímetros.
- **body_mass_g:** Masa corporal del pingüino en gramos.
- **sex:** Género del pingüino, con las categorías 'Male' (macho), 'Female' (hembra) o 'NaN' si la información no está disponible.

Con el objetivo de reducir el número de variables numéricas y explorar relaciones entre las características físicas de los pingüinos, así como entre las especies, realizar los siguientes apartados:

1. Calcular la matriz de correlaciones y su representación gráfica: ¿Cuáles son las variables más correlacionadas entre las características físicas de los pingüinos?
2. Realizar un análisis de componentes principales (PCA) sobre la matriz de correlaciones, calculando un número adecuado de componentes (máximo 4): Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes para representar eficientemente la variabilidad de las especies de pingüinos?
3. Realizar nuevamente el análisis de componentes principales sobre la matriz de correlaciones, esta vez indicando el número de componentes principales que hemos decidido retener. Sobre este análisis, contestar los siguientes apartados:
   a) Escribe numéricamente el proceso para el cálculo de la primera observación en la primera componente.
   b) Comentar los gráficos que representan las variables en los planos formados por las componentes: Intenta explicar lo que representa cada componente en términos de las características físicas de los pingüinos.
   c) Sobre los gráficos que representan las observaciones en los nuevos ejes: Teniendo en cuenta la posición de las especies de pingüinos en el gráfico, ¿cuáles destacan más en cada componente?

d) Comenta Relación entre las Componentes Principales y las Variables, las contribuciones de las Componentes Principales a la Variabilidad Explicada de las Variables Originales y la contribuciones de las Variables a las Componentes Principales.

Tras explorar las características físicas de los pingüinos mediante el Análisis de Componentes Principales (ACP), ahora nos centraremos en evaluar la estructura de los grupos o clústeres existentes en el conjunto de datos. Utilizaremos técnicas de análisis de clúster, tanto jerárquicas como no jerárquicas, para identificar posibles agrupaciones basadas en las características (numéricas) de los pingüinos. Los siguientes son los pasos y tareas a realizar:

1. **Calcular la Matriz de Distancias:** Inicialmente, calcular la matriz de distancias entre las observaciones, empleando una medida de distancia adecuada, como la distancia euclidia.

2. **Determinación del Número de Clústeres - Análisis Jerárquico:** Utilizar métodos de clustering jerárquico para explorar la estructura de los datos. Aplicar técnicas como el dendrograma para visualizar y determinar el número óptimo de clústeres. Aplicar los procemidientos que conozcas que sirvan para decidir el número de clusters. Discutir las decisiones tomadas en este proceso.

3. **Análisis de Clúster No Jerárquico:** Una vez decidido el número de clústeres, realizar un análisis de clúster no jerárquico (K-means) utilizando el número de grupos identificado en el paso anterior.

4. **Evaluación de la Calidad de las Agrupaciones:** Evaluar la calidad de los clústeres formados utilizando el índice de silueta.

5. **Variables suplementarias:** Representa los centroides de cada especie y cada isla en un gráfico que tenga como ejes las dos primeras componentes principales.

6. **Caracterización de los Clústeres:** Caracterizar cada clúster basándose en estadísticos descriptivos y comparar las propiedades distintivas entre ellos.

La puntuación de la tarea está dividida en 2 partes:

- Análisis de Componentes principales (5 puntos). Todos los apartados de esta parte tienen la misma puntuación.

- Análisis Cluster (5 puntos). Todos los apartados de esta parte tienen la misma puntuación.