ANÁLISIS DE COMPONENTES PRINCIPALES Y CLUSTERING
Dataset Palmer Penguins

Alumno: Gerson Castillo
DNI: 54768597T
Profesor: Pablo Arcadio
Fecha de Entrega: 18 de mayo de 2025

CONTENIDO

1. INTRODUCCIÓN

La caracterización de grupos naturales dentro de conjuntos de datos biológicos resulta fundamental para comprender la variabilidad intra-específica y las adaptaciones ecológicas. Entre las técnicas de aprendizaje no supervisado, el **Análisis de Componentes Principales (ACP)** y los métodos de **clustering** permiten, respectivamente, reducir la dimensionalidad de la información y detectar estructuras latentes en los datos .

El conjunto de datos **Palmer Penguins** ha ganado popularidad como alternativa pedagógica al clásico *Iris*, ya que recopila medidas morfológicas de tres especies de pingüinos (Adelie, Chinstrap y Gentoo) recolectadas en las islas Palmer, Antártida. Las variables incluyen el largo y la profundidad del pico, la longitud de la aleta y la masa corporal, junto con factores categóricos como la especie, la isla de procedencia y el sexo.

El presente informe tiene por finalidad aplicar ACP para identificar los ejes principales de variación morfológica y, sobre dichos componentes, ejecutar algoritmos de clustering jerárquico y **K-means** con el fin de segmentar a los individuos en grupos homogéneos. Finalmente, se discutirán los resultados en términos biológicos y se destacarán las limitaciones del enfoque.

2. MATERIALES Y MÉTODOS

2.1. Conjunto de datos

El dataset completo consta de 344 registros; sin embargo, tras eliminar los valores faltantes mediante *list-wise deletion*, se trabajó con **333 individuos**. Las variables numéricas utilizadas fueron: *bill_length_mm*, *bill_depth_mm*, *flipper_length_mm* y *body_mass_g*. Las variables categóricas (*species*, *island*, *sex*) se reservaron para la interpretación posterior.

2.2. Preprocesamiento

1. **Limpieza de datos:** Se descartaron filas con valores nulos para simplificar el flujo de análisis.
2. **Estandarización:** Las variables numéricas se escalaron con *StandardScaler* (media = 0, desviación = 1) para evitar sesgos por diferencias de escala.
3. **División de variables:** Se separaron las variables numéricas (input para ACP y clustering) de las categóricas (usadas solo para interpretación).
4.

2.3. Análisis de Componentes Principales (ACP)

Se aplicó ACP usando la librería *scikit-learn* y las funciones de visualización proporcionadas por **FuncionesMineria2.py**. El número de componentes retenidos se determinó mediante el criterio del codo en el **scree plot** y la regla de varianza acumulada ≥ 80 %.

2.4. Clustering

- **Clustering jerárquico**: Se empleó el método de Ward sobre las puntuaciones de los componentes principales y se examinó el dendrograma para sugerir un valor inicial de *k*.
- **K-means**: Se exploraron valores de *k* entre 2 y 10. El *k* óptimo se seleccionó combinando el gráfico del codo (WCSS) y la silueta media.

2.5. Evaluación y perfilado de grupos

Se generaron tablas de contingencia y estadísticas descriptivas por cluster. Además, se comparó la asignación de individuos entre el clustering jerárquico y K-means para evaluar la estabilidad de las agrupaciones.

3. FASE A – EXPLORACIÓN INICIAL

3.1. Exploración inicial (Fase A)

Código:

```
# Carga y vista inicial de datos
df = sns.load_dataset("penguins")
display(df.head())
```

Se examinó la estructura del conjunto de datos: **344 registros** y **7 variables**. ...

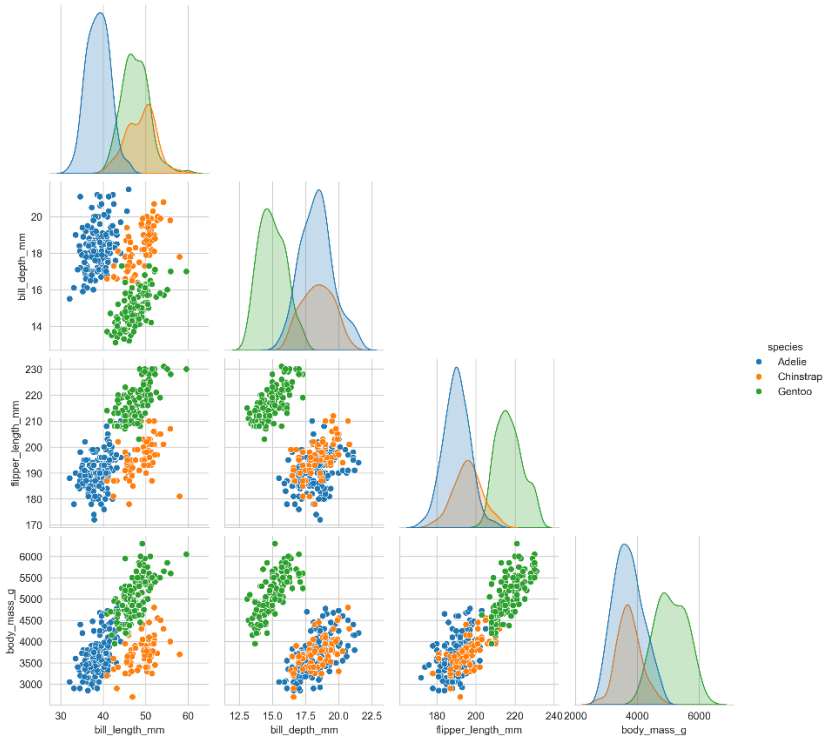
Se examinó la estructura del conjunto de datos: **344 registros** y **7 variables**. Se identificaron **2 valores faltantes** en cada variable morfológica y **11 en la variable *sex***.

Las estadísticas descriptivas de las variables numéricas muestran:

- **bill_length_mm**: media = 43.92 mm, σ = 5.46 mm, rango [32.1, 59.6] mm.
- **bill_depth_mm**: media = 17.15 mm, σ = 1.97 mm, rango [13.1, 21.5] mm.
- **flipper_length_mm**: media = 200.92 mm, σ = 14.06 mm, rango [172, 231] mm.
- **body_mass_g**: media = 4201.75 g, σ = 801.95 g, rango [2700, 6300] g.

Tras eliminar filas con valores faltantes (*list-wise deletion*), se trabajó con **333 individuos** (96 % de la muestra).

Matriz de diagramas de dispersión y densidad



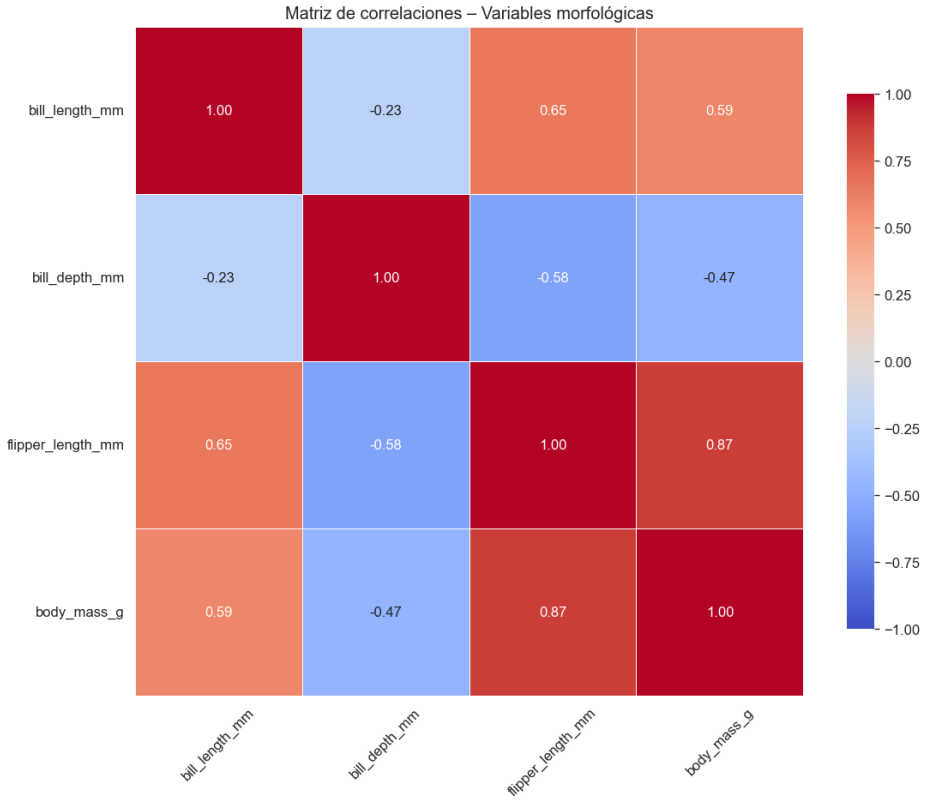
Observaciones clave:

- **Separación por especie:** Los individuos *Gentoo* (verde) se agrupan claramente hacia valores elevados de largo de pico, longitud de aleta y masa corporal.
- **Solapamiento parcial:** *Adelie* (azul) y *Chinstrap* (naranja) se solapan, aunque los *Chinstrap* tienden a presentar picos más largos a igual profundidad.

- **Relaciones lineales:** Se aprecia una marcada relación positiva entre *flipper_length_mm* y *bill_length_mm*, y entre ambas con *body_mass_g*.
- **Outliers:** Pocos puntos aislados sugieren escasa presencia de valores atípicos.

4. FASE B – ANÁLISIS DE CORRELACIONES

Heat-map de la matriz de correlaciones



Interpretación:

- *bill_length_mm* y *bill_depth_mm*: correlación débil negativa (r = −0.23).
- *bill_length_mm* y *flipper_length_mm*: correlación moderada positiva (r = 0.65).
- *bill_length_mm* y *body_mass_g*: correlación moderada positiva (r = 0.59).
- *bill_depth_mm* y *flipper_length_mm*: correlación moderada negativa (r = −0.58).
- *bill_depth_mm* y *body_mass_g*: correlación moderada negativa (r = −0.47).
- *flipper_length_mm* y *body_mass_g*: correlación fuerte positiva (r = 0.87).

Estas relaciones muestran redundancias y dependencias claras entre medidas de tamaño, lo que justifica el uso de ACP para sintetizar la información y mitigar la multicolinealidad.

5. FASE C – VARIABILIDAD EXPLICADA POR EL ACP

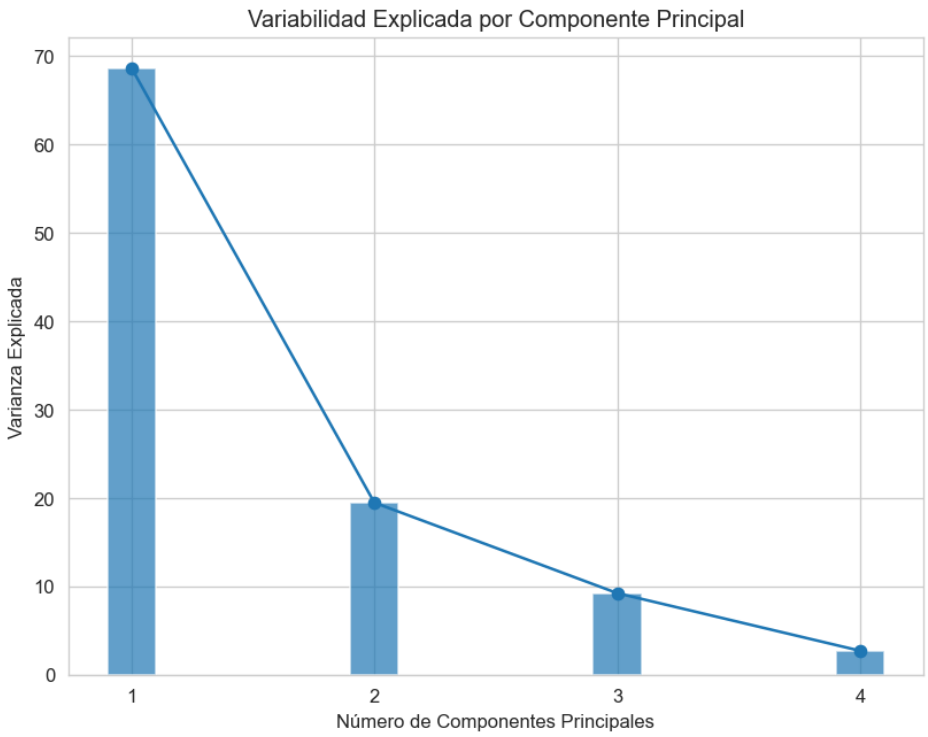
Código:

```
# 1. Variables numéricas
vars_num = ["bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g"]
notas = df_clean[vars_num]

# 2. Estandarización
notas_estandarizadas = pd.DataFrame(
    StandardScaler().fit_transform(notas),
    columns=[f"{col}_z" for col in vars_num],
    index=df_clean.index
)

# 3. PCA completo (todas las componentes)
pca_full = PCA(n_components=len(vars_num))
fit_full = pca_full.fit(notas_estandarizadas)

# 4. Autovalores y varianza explicada
autovalores    = fit_full.explained_variance_
var_explicada  = fit_full.explained_variance_ratio_ * 100
var_acumulada  = np.cumsum(var_explicada)

# 5. Tabla de resultados
tabla_pca = pd.DataFrame(
    {
        "Autovalores": autovalores,
        "Variabilidad Explicada (%)": var_explicada,
        "Variabilidad Acumulada (%)": var_acumulada
    },
    index=[f"Componente {i}" for i in range(1, fit_full.n_components_ + 1)]
)
print(tabla_pca)

# 6. Scree plot (función oficial)
plot_varianza_explicada(var_explicada, fit_full.n_components_)
```

Tabla 1. Autovalores y variabilidad explicada

	Autovalores	Variabilidad Explicada (%)	Variabilidad Acumulada (%)
Componente 1	2.753625	68.633893	68.633893
Componente 2	0.780461	19.452929	88.086822
Componente 3	0.369753	9.216063	97.302885
Componente 4	0.108210	2.697115	100.000000

Scree plot de varianza explicada



Interpretación:

- El "codo" aparece claramente en el componente 2, indicando que los dos primeros ejes resumen la mayor parte de la variabilidad del conjunto de datos.
- El **Componente 1**, con un autovalor de 2.7536, explica el 68.63 % de la variabilidad.
- El **Componente 2**, con un autovalor de 0.7805, añade un 19.45 % adicional.
- En conjunto, los dos primeros componentes acumulan el 88.09 % de la varianza total, superando el umbral del 80 % y validando la decisión de reducir la dimensión a 2 ejes.

La retención de dos componentes permite simplificar el análisis posterior sin perder información crítica. Estos ejes sintetizan la variabilidad biológica ligadas a las dimensiones morfológicas, facilitando la interpretación y el posterior clustering de los individuos.

6. FASE D – PCA DETALLADO E INTERPRETACIÓN

Se realiza un nuevo Análisis de Componentes Principales (PCA) sobre los datos estandarizados, esta vez restringido a las dos componentes principales seleccionadas en la fase anterior por explicar conjuntamente más del 88 % de la variabilidad total. A partir de esta nueva descomposición, se analizan las cargas, los gráficos, y se propone la construcción de un índice sintético morfológico.

6.1. Representación de las variables en el plano de componentes

Para interpretar las variables en el nuevo espacio de componentes, se calcularon las cargas o "loadings", es decir, los coeficientes que relacionan cada variable original con las componentes principales:

Código:

```
# 2. Cálculo de loadings (cargas) de cada variable
df_loadings = pd.DataFrame(
    pca_opt.components_.T,
    index=notas_estandarizadas.columns,
    columns=['PC1', 'PC2']
)
print("Cargas de las variables en PC1 y PC2:\n", df_loadings, "\n")
```

Resultados:

Variable	PC1	PC2
bill_length_mm_z	0.4538	0.6002
bill_depth_mm_z	-0.3990	0.7962
flipper_length_mm_z	0.5768	0.0058
body_mass_g_z	0.5497	0.0765

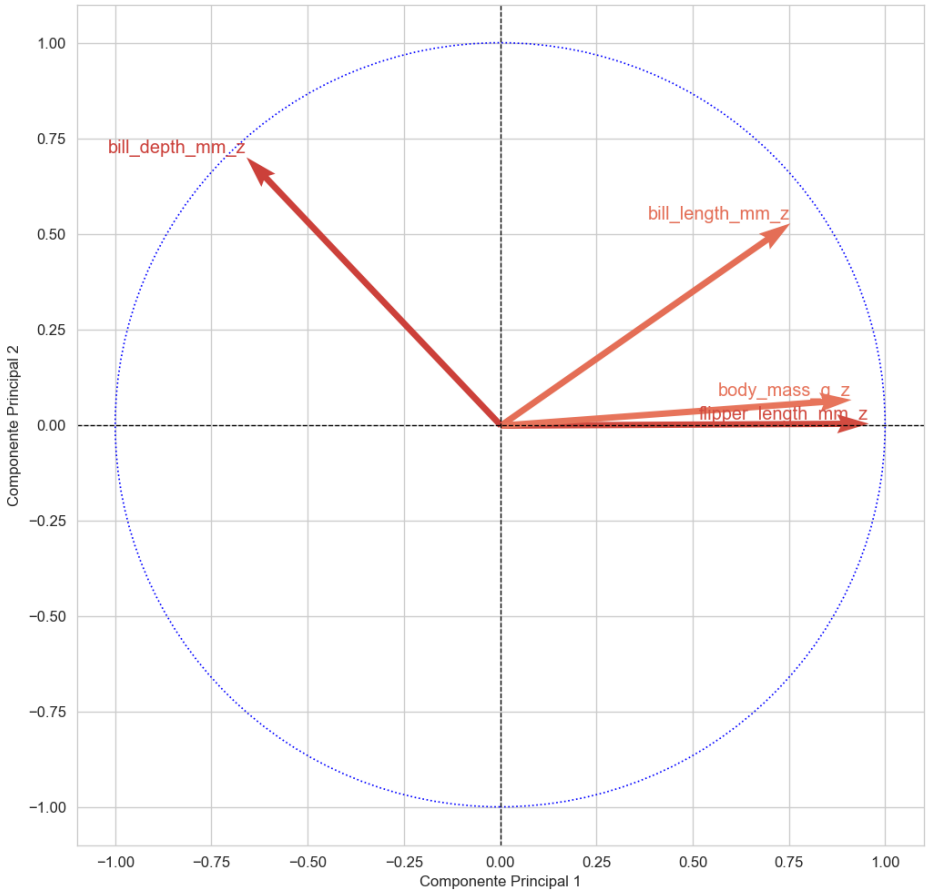
Estas cargas muestran que:

- La **Componente 1 (PC1)** recoge fundamentalmente la variación asociada al tamaño corporal general, representado por flipper_length_mm_z y body_mass_g_z (cargas > 0.54).
- La **Componente 2 (PC2)** captura aspectos relacionados con la forma del pico, siendo bill_depth_mm_z la variable más relevante (0.80), junto con bill_length_mm_z.

La orientación de estas variables en el gráfico de círculo de correlaciones generado mediante:

Código:

```
# 6. Círculo de correlaciones
plot_corr_cos(2, correlaciones)
```



Interpretación del círculo de correlaciones

En el gráfico, cada variable se representa como un vector dentro de una circunferencia de radio unitario. Las variables flipper_length_mm_z y body_mass_g_z se proyectan fuertemente sobre el eje PC1, lo que confirma su alta correlación con esta componente. Por otro lado, bill_depth_mm_z se alinea casi por completo con PC2, indicando que su variación no está relacionada con el tamaño corporal sino con la geometría del pico.

La longitud de cada flecha indica el grado en que una variable está bien representada en el plano. Las más largas (como flipper_length y bill_depth) son las mejor explicadas por las dos componentes retenidas. Las variables cercanas entre sí (como flipper_length y body_mass) están positivamente correlacionadas, mientras que vectores perpendiculares indican independencia.

6.2. Distribución de las especies en el nuevo espacio PCA

Para estudiar cómo se posicionan las especies en el espacio reducido, se calcularon los scores individuales y las medias por especie:

Especie	PC1	PC2
Adelie	−1.460	−0.142
Chinstrap	−0.389	+0.993
Gentoo	+2.013	−0.394

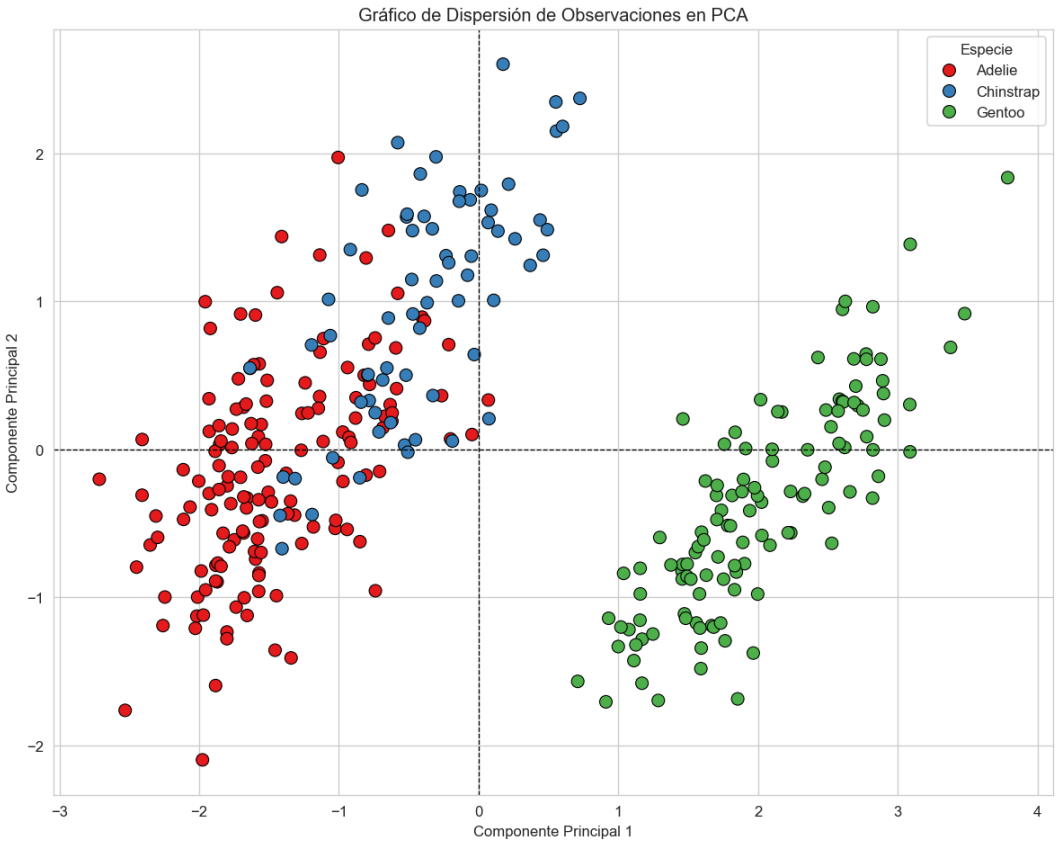
Interpretación:

- **Gentoo** presenta valores muy altos en PC1 y bajos en PC2. Esto implica mayor tamaño corporal y menor profundidad de pico.
- **Chinstrap** destaca en PC2, indicando una forma de pico más largo y menos profundo.
- **Adelie** aparece en la zona negativa de ambas componentes, con menor masa corporal y un pico más compacto.

Estos resultados se visualizaron con el siguiente comando:

Código:

```
df_scores = pd.DataFrame(X_pca, columns=['PC1', 'PC2'], index=df_clean.index)
df_scores['species'] = df_clean['species'].values


def plot_pca_scatter_colored(df_scores, title="Gráfico de Dispersión de Observaciones en PCA"):
    """
    Muestra un scatter plot de observaciones proyectadas en las dos primeras componentes
principales,
    coloreando los puntos según la especie.

    Parámetros:
    - df_scores: DataFrame con columnas ['PC1', 'PC2', 'species'].
    """
    plt.figure(figsize=(10, 8))
    sns.scatterplot(
        data=df_scores,
        x='PC1',
        y='PC2',
        hue='species',
        palette='Set1',  # Puedes cambiar la paleta: 'Set2', 'Set3', 'Dark2', etc.
        edgecolor='black',
        s=70
    )
    plt.axhline(0, linestyle='--', color='black', linewidth=0.8)
    plt.axvline(0, linestyle='--', color='black', linewidth=0.8)
    plt.title(title)
    plt.xlabel("Componente Principal 1")
    plt.ylabel("Componente Principal 2")
    plt.legend(title='Especie')
    plt.grid(True)
    plt.tight_layout()
    plt.show()

plot_pca_scatter_colored(df_scores)
```



El gráfico de dispersión coloreado por especie revela una **segmentación clara** entre los tres grupos:

- **Gentoo**: Se concentra en el extremo derecho (valores altos en PC1), lo que indica **mayor masa corporal y aletas más largas**, coherente con su morfología robusta.
- **Chinstrap**: Se ubica mayoritariamente en la parte superior central del gráfico (valores intermedios en PC1, pero elevados en PC2). Esta especie destaca por un **pico más largo y menos profundo**.
- **Adelie**: Se agrupa en la parte izquierda (valores negativos en PC1), indicando **menor tamaño corporal**. También ocupa valores moderados en PC2, reflejando **un pico más profundo y corto**.

Este gráfico es especialmente útil porque muestra que las dos dimensiones extraídas por el ACP permiten una **discriminación eficaz entre las especies**, lo cual valida su capacidad para reducir dimensionalidad sin perder información relevante.

6.3. Construcción de un índice morfológico sintético

A partir de los resultados obtenidos en PC1, se construye un índice morfológico continuo que sintetiza las dimensiones físicas relevantes de cada individuo. Para asegurar que todos los valores del índice sean estrictamente positivos, se aplica una transformación exponencial a la combinación lineal obtenida en PC1.

El índice resultante se expresa como:

$$\text{Índice}_{\text{PC1+}} = \exp(0.4538 \cdot bill\_length\_mm\_z - 0.3990 \cdot bill\_depth\_mm\_z + 0.5768 \cdot flipper\_length\_mm\_z + 0.5497 \cdot body\_mass\_g\_z)$$

Resultados promedio por especie:

- **Adelie**: 0.273
- **Chinstrap**: 0.784
- **Gentoo**: 9.190

Estos valores confirman que el índice refleja adecuadamente las diferencias morfológicas entre especies, con Gentoo destacándose como la especie de mayor tamaño estructural. El índice resultante puede utilizarse como variable compuesta en análisis posteriores de agrupamiento o clasificación.

7. FASE E – CLUSTERING JERÁRQUICO

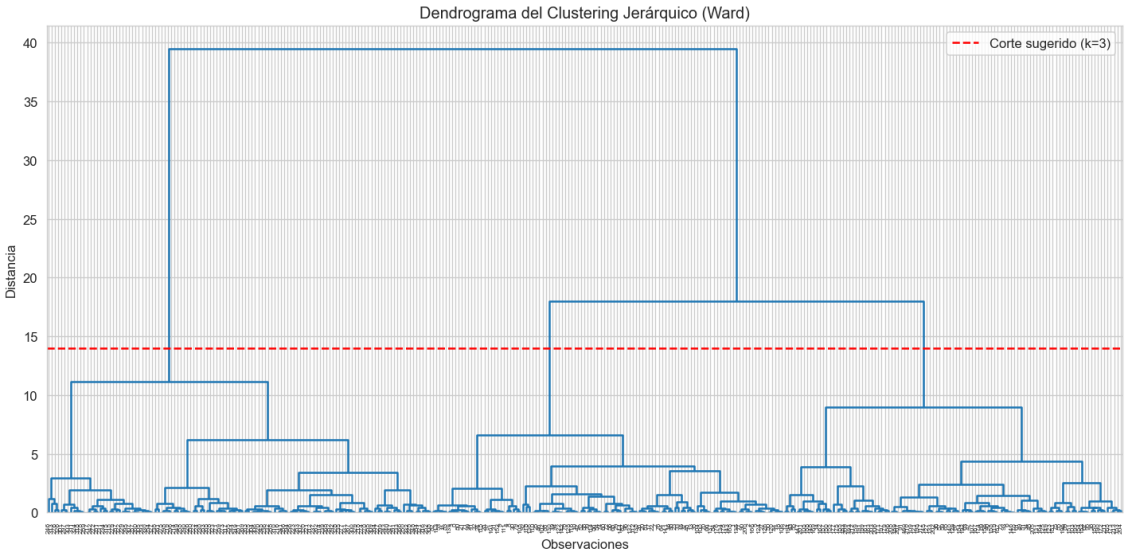
En esta fase se aplica un algoritmo de agrupamiento jerárquico con el método de Ward sobre las observaciones previamente transformadas por el Análisis de Componentes Principales (PCA). Se utiliza únicamente la representación en las dos primeras componentes principales (PC1 y PC2), las cuales explican más del 88 % de la variabilidad original.

El análisis se realiza en dos etapas: primero, se calcula la matriz de distancias euclídeas entre individuos en el espacio reducido; luego, se construye el dendrograma para visualizar las sucesivas fusiones de clústeres y proponer un valor razonable de k.

Código:

```
# Usamos solo las componentes PC1 y PC2 para el clustering
X_cluster = df_scores[['PC1', 'PC2']].values

# Cálculo del linkage jerárquico con método 'ward'
Z = linkage(X_cluster, method='ward')

# Dendrograma
plt.figure(figsize=(12, 6))
dendrogram(Z, color_threshold=0, leaf_rotation=90)
plt.axhline(y=14, color='red', linestyle='--', label='Corte sugerido (k=3)')
plt.title("Dendrograma del Clustering Jerárquico (Ward)")
plt.xlabel("Observaciones")
plt.ylabel("Distancia")
plt.legend()
plt.tight_layout()
plt.show()
```

Figura del dendrograma



Dendrograma jerárquico con método de Ward sobre las componentes principales PC1 y PC2. El gráfico muestra la estructura de agrupamiento de los individuos en función de sus coordenadas en el espacio PCA. La línea discontinua roja indica el corte sugerido a una distancia ≈ 14, a partir del cual se identifican tres grupos principales (k = 3).

Interpretación del dendrograma

El dendrograma muestra un salto abrupto en la altura de fusión entre los tres grupos más grandes, en torno a una **distancia de corte de 14**. De acuerdo con el criterio del "último gran salto", este punto representa una transición significativa en la heterogeneidad entre clústeres. Por tanto, se propone un corte del dendrograma en dicha altura, lo que permite identificar **k = 3 grupos** como una partición coherente.

Este número de clústeres no solo está justificado por la estructura visual del dendrograma, sino que también coincide con el número de especies conocidas en el conjunto de datos. Esta hipótesis será evaluada y contrastada en la siguiente fase, mediante el algoritmo k-means y medidas de validación como el gráfico del codo y el coeficiente de silueta.

8. FASE F – K-MEANS Y VALIDACIÓN DEL NÚMERO DE CLÚSTERES

En esta fase se aplica el algoritmo de K-means sobre los datos proyectados en el espacio reducido generado por el Análisis de Componentes Principales (PC1 y PC2). A diferencia del enfoque jerárquico, que construye una estructura de agrupamiento de forma ascendente, el método K-means requiere definir previamente el número de clústeres, por lo que se evalúa una gama de valores de k comprendida entre 2 y 10.

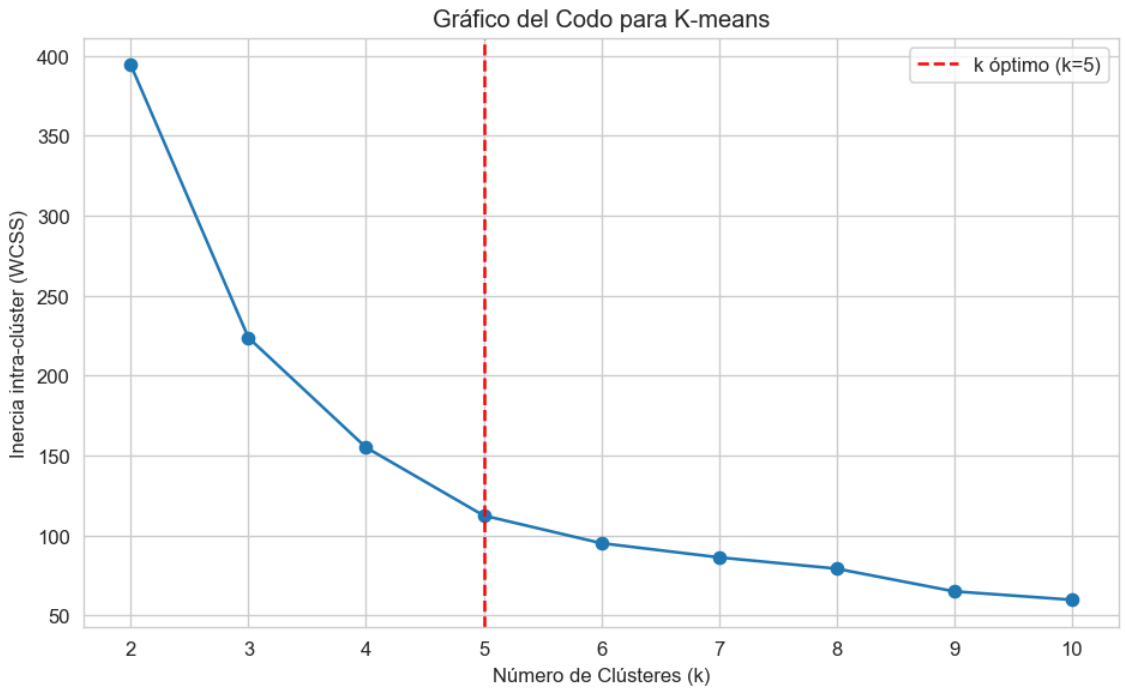
Para identificar el número óptimo de grupos, se emplean dos criterios clásicos de validación:

- El **gráfico del codo**, que analiza la inercia intra-clúster (Within-Cluster Sum of Squares, WCSS) y permite detectar el punto donde agregar más clústeres no mejora significativamente la compactación.
- El **coeficiente de silueta media**, que evalúa la cohesión y separación entre grupos.

Código:

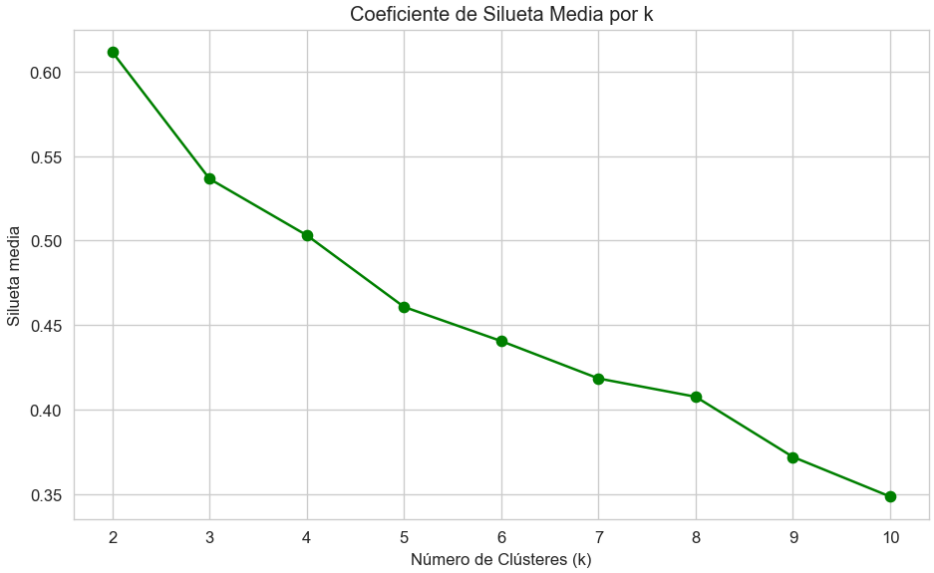
```
# Clustering K-means
# Usamos las componentes PC1 y PC2 para el clustering K-means
# Aseguramos que las columnas estén en el orden correcto

X_kmeans = df_scores[['PC1', 'PC2']].values

inertias = []
silhouette_scores = []
ks = range(2, 11)

for k in ks:
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(X_kmeans)
    inertias.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(X_kmeans, labels))
```

Gráfico del codo para K-means.



La inercia intra-clúster (WCSS) decrece rápidamente entre k = 2 y k = 5. A partir de **k = 5**, las reducciones se vuelven gradualmente menos pronunciadas. Este comportamiento sugiere que **k = 5 es el punto óptimo**, ya que representa el límite a partir del cual las ganancias en compactación son marginales. Esta interpretación es coherente con el criterio del "codo" que busca el valor de k en el que la mejora deja de ser significativa.

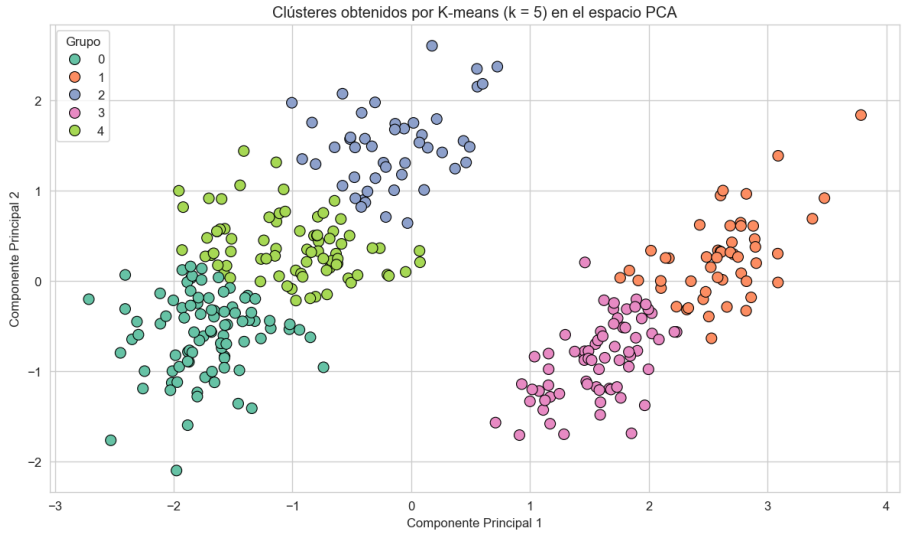
Coeficiente de silueta media por número de clústeres.



El coeficiente de silueta media alcanza su máximo en **k = 2 (≈ 0.61)**, pero disminuye de forma progresiva con el aumento de k. Para **k = 5**, el valor de la silueta sigue siendo aceptable (≈ 0.46), lo que indica una partición razonablemente coherente a pesar de no ser la más "pura". Este valor es superior al observado para k = 6 en adelante.

Aplicación de K-means con k = 5

Tras validar que k = 5 es una elección razonable, se ejecuta el algoritmo K-means con ese valor. Las observaciones se agrupan en cinco clústeres, cuya distribución en el plano PCA se muestra a continuación:



Distribución de clústeres obtenidos por K-means (k = 5) en el espacio PCA.
Cada punto representa un individuo, coloreado según su asignación de grupo. Se observa una segmentación clara, con grupos diferenciados que ocupan regiones específicas del espacio PCA, particularmente a lo largo del eje PC1, que está asociado al tamaño corporal.

Conclusión

La evaluación conjunta del gráfico del codo y del coeficiente de silueta media permite concluir que **k = 5** representa un número óptimo de clústeres para este conjunto de datos. Esta solución ofrece un buen equilibrio entre compacidad y separación de grupos, y será utilizada en la siguiente fase para caracterizar los perfiles morfológicos de cada grupo e interpretar su posible relación con las especies reales.

9. FASE G – COMPARACIÓN Y PERFILADO DE GRUPOS

En esta fase se comparan los resultados obtenidos mediante los algoritmos de agrupamiento jerárquico y K-means, ambos aplicados sobre el espacio PCA (PC1 y PC2). El objetivo es contrastar las asignaciones de grupo, identificar similitudes y diferencias, y caracterizar morfológicamente los clústeres generados por K-means, discutiendo su posible interpretación biológica en el contexto de las especies de pingüinos.

9.1 Comparación entre agrupamiento jerárquico y K-means

Ambos métodos partieron del mismo espacio proyectado por PCA, pero generaron agrupamientos distintos debido a sus diferencias metodológicas:

- El **agrupamiento jerárquico** (método de Ward) propuso una segmentación en **tres grupos**, basándose en las fusiones jerárquicas de los individuos.
- El **K-means**, tras analizar la inercia intra-clúster y la silueta media, permitió adoptar una segmentación más detallada con **cinco grupos (k = 5)**.

El siguiente código muestra la tabla cruzada entre los grupos de K-means y las especies reales, normalizada por fila:

Código:

```
pd.crosstab(df_scores['cluster_kmeans_5'], df_scores['species'], normalize='index').round(2)
```

Clúster	Adelie	Chinstrap	Gentoo
0	0.94	0.06	0.00
1	0.00	0.00	1.00
2	0.15	0.85	0.00
3	0.00	0.00	1.00
4	0.71	0.29	0.00

Se observa que:

- Los clústeres **1 y 3** contienen exclusivamente individuos de la especie **Gentoo**, lo que sugiere subgrupos internos con diferencias morfológicas dentro de la misma especie.
- El grupo **0** representa principalmente a **Adelie**, con morfología más compacta.
- El grupo **2** se asocia claramente con **Chinstrap**, caracterizado por picos largos y profundos.
- El grupo **4** muestra una mezcla sustancial entre **Adelie (71 %)** y **Chinstrap (29 %)**, sin presencia de Gentoo.

9.2 Interpretación morfológica de los grupos

Se calcularon las medias y desviaciones estándar de las variables morfológicas estandarizadas por grupo:

Código:

```
df_clustered = df_scores.join(notas_estandarizadas)
group_summary = df_clustered.groupby('cluster_kmeans_5')[[
    'bill_length_mm_z', 'bill_depth_mm_z', 'flipper_length_mm_z', 'body_mass_g_z'
]].agg(['mean', 'std']).round(2)
group_summary
```

Clúster	bill_length	bill_depth	flipper_length	body_mass
0	−1.17 ± 0.41	0.24 ± 0.46	−0.98 ± 0.38	−0.98 ± 0.37
1	1.15 ± 0.43	−0.71 ± 0.38	1.55 ± 0.39	1.67 ± 0.36
2	1.10 ± 0.59	1.06 ± 0.44	−0.15 ± 0.43	−0.25 ± 0.48
3	0.30 ± 0.34	−1.39 ± 0.37	0.88 ± 0.29	0.69 ± 0.40
4	−0.33 ± 0.59	0.77 ± 0.59	−0.56 ± 0.42	−0.40 ± 0.49

El grupo 1 representa pingüinos con mayor masa corporal y aletas más largas (correspondiente a Gentoo), mientras que el grupo 0 agrupa individuos pequeños y robustos (Adelie). El grupo 2 muestra un pico largo y profundo pero cuerpo más liviano (Chinstrap). **El grupo 3** refleja una subpoblación intermedia de Gentoo.

El grupo 4, sin embargo, exhibe una combinación morfológica intermedia entre **Adelie y Chinstrap**, lo que se manifiesta tanto en las proporciones de especie como en los valores morfométricos. Esta distribución sugiere que el grupo podría representar:

- Una **subpoblación morfológicamente híbrida**, con rasgos intermedios no claramente alineados con una sola especie.
- Posiblemente, un **evento de hibridación natural** o una diferenciación incipiente que podría derivar en una nueva forma evolutiva o incluso en una futura especie si se mantienen las diferencias.

Esta hipótesis plantea una dirección interesante para estudios de biología evolutiva y genética de poblaciones, ya que los métodos de clustering no supervisado han permitido aislar un grupo morfológicamente coherente pero taxonómicamente mixto.

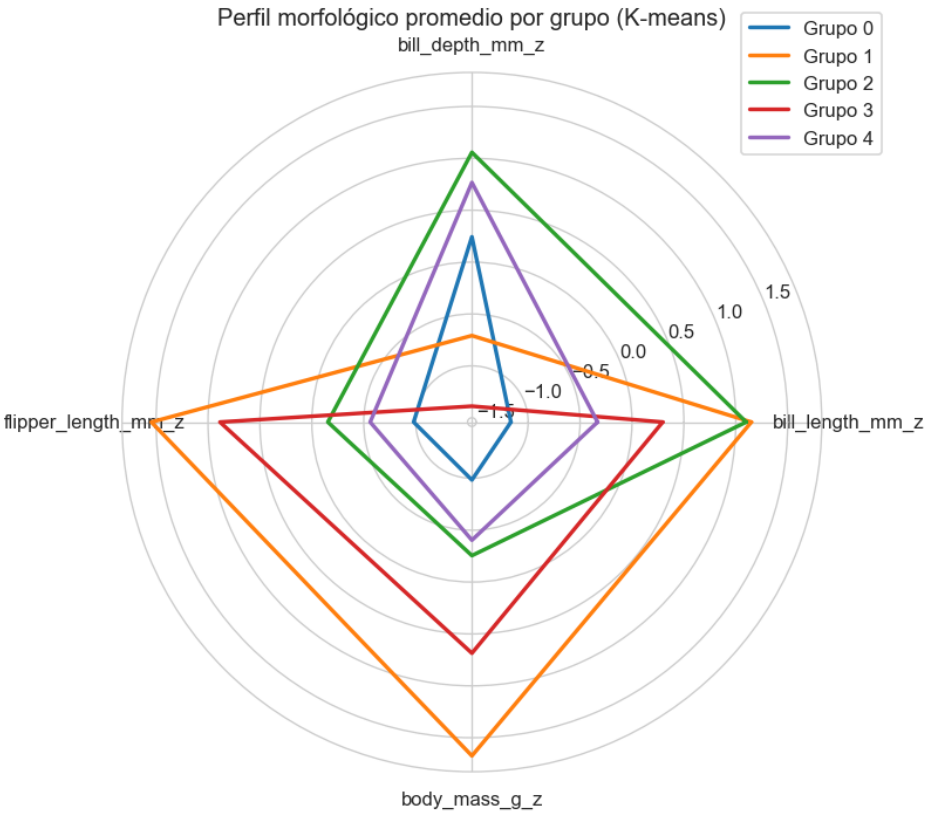
9.3 Visualización del perfil de grupo

Para facilitar la comparación visual, se generó un gráfico radar con el perfil promedio de cada grupo:

Código:

```
variables = ['bill_length_mm_z', 'bill_depth_mm_z', 'flipper_length_mm_z', 'body_mass_g_z']
group_means = df_clustered.groupby('cluster_kmeans_5')[variables].mean()

angles = np.linspace(0, 2 * np.pi, len(variables), endpoint=False).tolist()
angles += angles[:1]

plt.figure(figsize=(8, 6))
for i, row in group_means.iterrows():
    values = row.tolist() + [row.tolist()[0]]
    plt.polar(angles, values, label=f'Grupo {i}', linewidth=2)

plt.xticks(angles[:-1], variables)
plt.title("Perfil morfológico promedio por grupo (K-means)")
plt.legend(loc='upper right', bbox_to_anchor=(1.1, 1.1))
plt.tight_layout()
plt.show()
```

Visualización del perfil de grupos con gráfico radar



El gráfico radar permite visualizar con claridad las diferencias de escala entre grupos. El grupo 1 domina en tamaño y masa, el grupo 2 en proporción del pico, mientras que el grupo 4 ocupa una posición intermedia en casi todas las dimensiones.

Conclusión

El modelo K-means con k = 5 ha permitido identificar no solo patrones que coinciden con las especies biológicas conocidas, sino también **subestructuras internas** dentro de especies como Gentoo, y **zonas de transición morfológica** como el grupo 4. Este último es particularmente interesante, ya que podría representar una **población intermedia evolutivamente relevante**, con implicaciones para el estudio de procesos de diferenciación o hibridación.

Esta fase confirma que los métodos de agrupamiento no supervisado aplicados sobre espacios reducidos mediante ACP permiten detectar patrones morfológicos complejos, más allá de las clasificaciones taxonómicas predefinidas.

10. FASE H – CONCLUSIONES Y LIMITACIONES

Conclusiones del análisis

El objetivo principal de este trabajo fue explorar la estructura latente del conjunto de datos Palmer Penguins a partir de sus variables morfológicas, empleando técnicas de análisis multivariante. Para ello, se aplicó un enfoque combinado de reducción de dimensionalidad (Análisis de Componentes Principales) y agrupamiento no supervisado (clustering jerárquico y particional).

El Análisis de Componentes Principales permitió representar adecuadamente la variabilidad morfológica en un espacio bidimensional, reteniendo más del 88 % de la varianza total con solo dos componentes. La primera componente sintetiza el tamaño corporal general, mientras que la segunda captura la forma del pico.

Sobre este espacio reducido, se aplicaron dos estrategias de agrupamiento:

- El **agrupamiento jerárquico** sugirió un número óptimo de **k = 3** clústeres, identificado visualmente en el dendrograma mediante el criterio del último gran salto. Esta partición general corresponde de forma aproximada a las tres especies reales del conjunto: Adelie, Chinstrap y Gentoo.
- El **algoritmo K-means**, validado con el gráfico del codo y la silueta media, permitió adoptar una solución más detallada con **k = 5** grupos. Esta elección aportó una visión más rica de la diversidad morfológica, revelando subestructuras internas dentro de las especies y zonas de transición entre ellas.

Entre los hallazgos más significativos se destaca el **grupo 4**, compuesto mayoritariamente por individuos de las especies Adelie (71 %) y Chinstrap (29 %). Este grupo presenta un perfil morfológico intermedio, tanto en longitud y profundidad del pico como en masa corporal. Su posición ambigua en el espacio PCA, junto con su composición mixta, sugiere la posibilidad de:

- Una **población híbrida** surgida de la interacción entre Adelie y Chinstrap.
- Un **proceso de diferenciación ecológica** que podría, con el tiempo, dar lugar a una nueva especie morfológica o funcional.

Este resultado plantea preguntas interesantes sobre la estructura genética y la ecología reproductiva de estas especies, y justifica la necesidad de estudios adicionales que integren datos genéticos, de comportamiento y distribución geográfica.

En conjunto, el análisis confirma que las técnicas de clustering aplicadas sobre componentes principales son herramientas eficaces para **identificar patrones emergentes en datos multivariados**, especialmente en contextos donde las categorías biológicas pueden ser insuficientes para explicar toda la variabilidad observada.

Limitaciones y desafíos del proceso de agrupamiento

Pese a los resultados obtenidos, es importante reconocer varias limitaciones metodológicas y técnicas que pueden haber influido en el análisis:

1. **Elección del número de clústeres en K-means**:
   El algoritmo requiere definir k previamente. Aunque se aplicaron criterios estándar (codo y silueta), la elección final conserva un componente subjetivo.
2. **Supuestos del algoritmo K-means**:
   Este método asume que los clústeres son esféricos, de tamaño comparable y distribuidos homogéneamente, lo cual puede no ajustarse a la morfología real de los pingüinos.
3. **Reducción de dimensionalidad con PCA**:
   Aunque muy útil para visualizar la estructura de los datos, el ACP proyecta en un espacio lineal y puede omitir relaciones no lineales o efectos de variables categóricas como la isla o el sexo.
4. **Validación limitada**:
   La única validación externa disponible fueron las etiquetas de especie. No se cuenta con datos genéticos, ecológicos o conductuales para confirmar la existencia de subespecies o híbridos reales.
5. **Pequeño tamaño muestral y valores faltantes**:
   El conjunto de datos tiene un tamaño reducido y requirió eliminar casos con datos incompletos, lo cual puede haber afectado la representatividad y precisión de los clústeres formados.

Recomendaciones para análisis futuros

- Incluir variables adicionales como **isla, sexo, dieta o comportamiento reproductivo**, que podrían mejorar la interpretación de los grupos.
- Integrar análisis **genético o geográfico** para verificar si los grupos morfológicos identificados corresponden a unidades biológicas reales.
- Replicar el análisis con un conjunto de datos ampliado para evaluar la robustez y estabilidad de los grupos formados.