



UNIVERSIDAD
COMPLUTENSE
MADRID

*MINERÍA DE DATOS Y
MODELIZACIÓN PREDICTIVA*

Kristian Sthefan Cortés Prieto

Introducción

En el mundo de la ciencia de datos, las técnicas de Análisis de Componentes Principales (ACP) y Clustering son fundamentales para descubrir patrones y relaciones ocultas en conjuntos de datos complejos. Este trabajo, solicitado por la sección de datos de National Geographic, tiene como objetivo principal aplicar estas técnicas al conjunto de datos penguins de la librería seaborn de Python.

El conjunto de datos penguins es una colección que comprende información detallada sobre diversas especies de pingüinos, recolectada de varias islas. Este conjunto incluye variables como la longitud del pico, profundidad del pico, longitud de la aleta, masa corporal y género de los pingüinos. El análisis de estos datos no solo revela aspectos interesantes sobre las características físicas de estas aves, sino que también proporciona insights valiosos sobre sus patrones de vida y adaptación ambiental.

En este trabajo, se enfoca en emplear ACP para reducir la dimensionalidad del conjunto de datos, facilitando así la visualización y comprensión de las relaciones entre las variables. Posteriormente, se aplicará técnicas de clustering para agrupar a los pingüinos en categorías significativas basadas en sus características físicas. Este enfoque no solo demuestra la aplicación práctica de estas técnicas en el análisis de datos biológicos, sino que también destaca su potencial para generar conocimientos relevantes en campos como la biología y la conservación de especies.

Desarrollo

1. Matriz de correlaciones y su representación gráfica

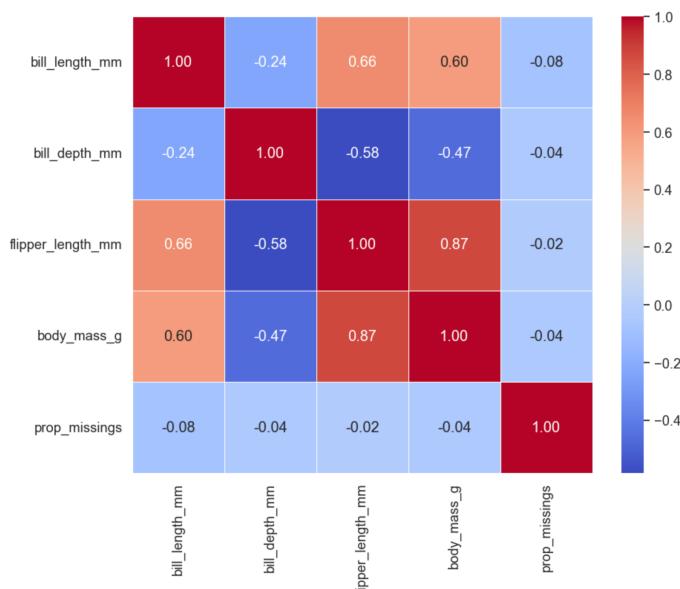


Imagen 1: Matriz de correlaciones

La imagen 1 muestra claramente las correlaciones entre diversas variables. La más destacada es la relación entre el tamaño de la aleta y la masa corporal del pingüino, con un coeficiente de correlación significativo de 0.87. Otras correlaciones notables incluyen la masa muscular y el tamaño del pico, con un coeficiente de 0.60, y el tamaño de la aleta correlacionado tanto con el tamaño del pico con un coeficiente de 0.66.

De igual forma se observa que la profundidad del pico presenta una correlación inversa significativa con el tamaño de la aleta, evidenciada por un coeficiente de correlación de -0.58. Esto implica que, a medida que aumenta la profundidad del pico, tiende a disminuir el tamaño de la aleta, y viceversa. De manera similar, la profundidad del pico también muestra una correlación inversa con la masa muscular del pingüino, con un coeficiente de -0.47, indicando que un aumento en la profundidad del pico generalmente coincide con una reducción en la masa muscular.

2. Análisis de componentes principales (PCA)

	Autovalores	Variabilidad Explicada	Variabilidad Acumulada
Componente 1	2.761831	68.843878	68.843878
Componente 2	0.774782	19.312919	88.156797
Componente 3	0.366307	9.130898	97.287695
Componente 4	0.108810	2.712305	100.000000

Imagen 2: PCA 4 componentes

Con la matriz de correlaciones se procedió a realizar un análisis de componentes principales utilizando el máximo posible de componentes, que en este caso es de cuatro. Este proceso permitió calcular la variabilidad explicada por cada componente. El objetivo era determinar el número óptimo de componentes para evitar el sobreprocesamiento de los datos y facilitar una interpretación más sencilla.

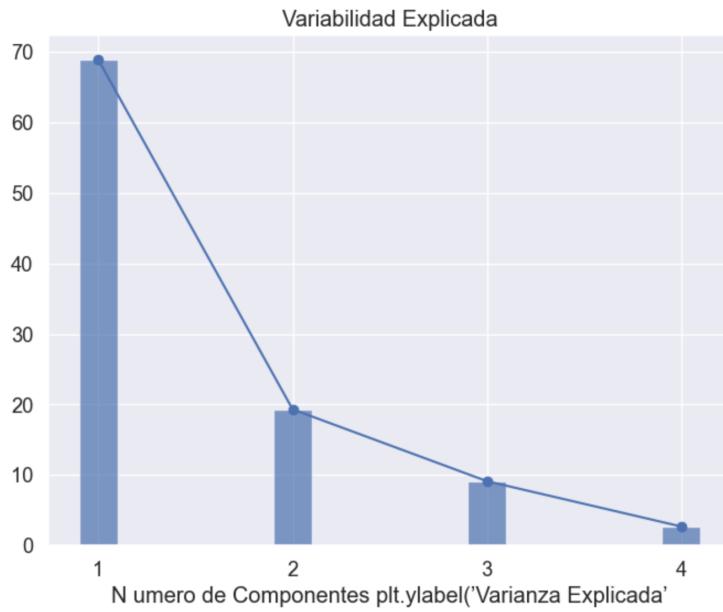


Imagen 3: Grafica del PCA 4 componentes

Como se puede observar en las figuras 2 y 3 al seleccionar únicamente dos componentes principales, se logra cubrir un significativo 88.15% de la variabilidad total. Por esta razón, se decidió proceder con solo dos componentes para el análisis subsiguiente. Esta elección permite una explicación más clara y eficiente de los datos, manteniendo al mismo tiempo una alta proporción de la información original.

3. PCA con componentes seleccionados

	Autovector 1	Autovector 2
bill_length_mm_z	0.455250	0.597031
bill_depth_mm_z	-0.400335	0.797767
flipper_length_mm_z	0.576013	0.002282
body_mass_g_z	0.548350	0.084363

Imagen 4: Autovectores 2 componentes

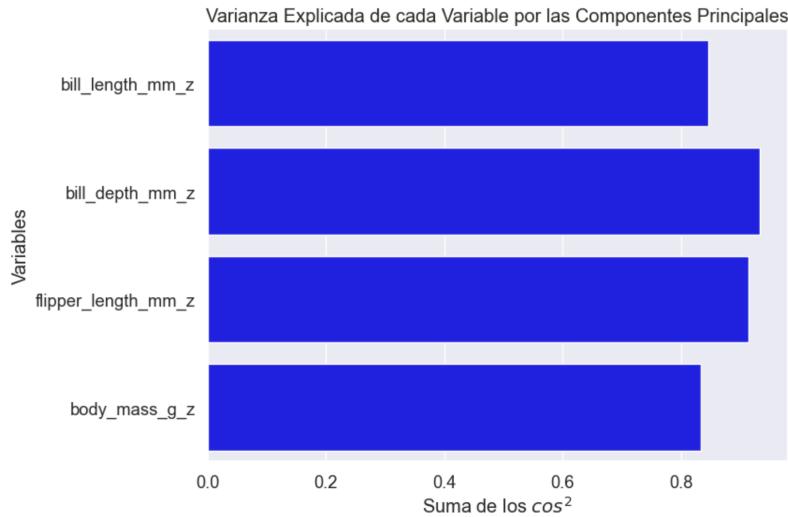


Imagen 5: Gráfico de barras para la suma de los cos2.

Las imágenes 4 y 5 nos ofrecen una visión detallada sobre la calidad de la representación en el análisis. Estas gráficas son particularmente útiles para evaluar hasta qué punto cada variable está adecuadamente representada en los componentes seleccionados. Se establece un umbral de calidad, donde un valor superior a 0.8 en estas imágenes indica una representación efectiva y precisa de la variable.

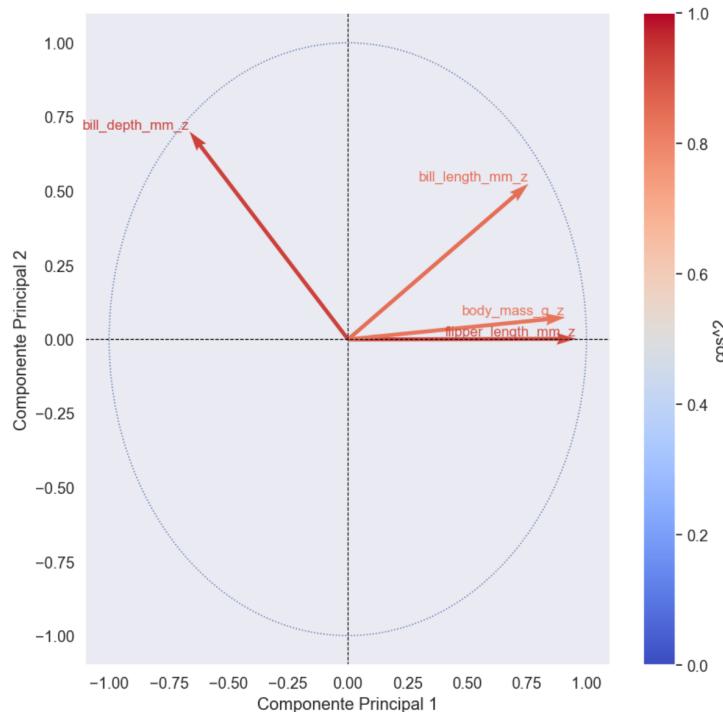


Imagen 6: Correlaciones entre Variables y Componentes Principales

El gráfico presentado en la imagen 6 y 7 se muestra la correlación entre diversos atributos, así como su influencia en dos componentes principales. En el gráfico 6 se observa una clara proximidad entre el tamaño de la aleta y la masa muscular, lo que indica una fuerte relación entre estas dos características, tal como se identificó previamente en la imagen 1. Asimismo, se destaca que estas variables afectan predominantemente al Componente 1, ejerciendo una influencia mínima sobre el Componente 2.

Además, el gráfico muestra que tanto la profundidad como el tamaño del pico tienen un impacto moderado en ambos componentes (1 y 2). Se aprecia que el tamaño del pico está directamente correlacionado con la masa muscular y el tamaño de la aleta. Por otro lado, la profundidad del pico muestra una correlación inversa con estas últimas, lo que sugiere una relación inversa significativa entre estas medidas.

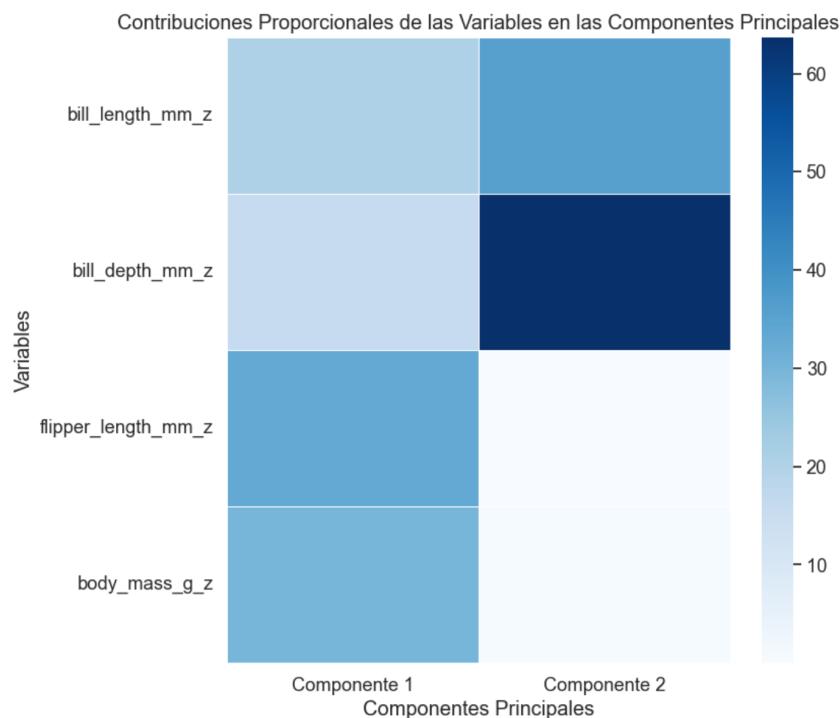


Imagen 7: Contribuciones de las Variables en las Componentes Principales

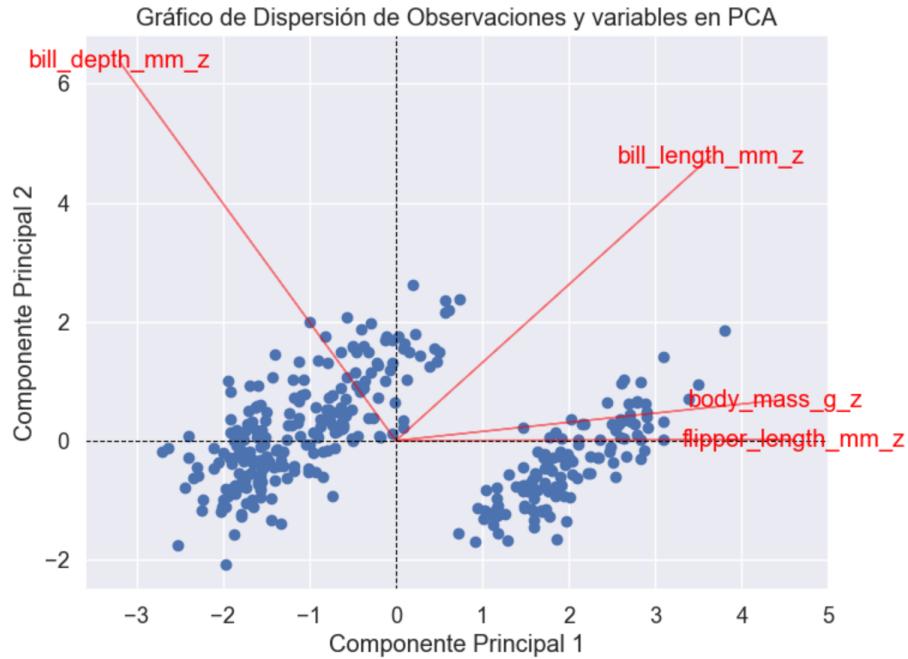


Imagen 8: Dispersión de los individuos en los nuevos ejes.

La imagen 8 da una perspectiva clara de cómo los datos de los pingüinos se relacionan con los nuevos ejes. Resulta evidente que aquellos pingüinos cuyos puntos de datos están más cerca de los ejes y más alejados del centro del gráfico tienden a tener un mayor desarrollo en ciertos atributos. Esta observación se hace más visible con el apoyo del gráfico en la imagen 9.

Por ejemplo, los pingüinos de la especie Gentoo se destacan por tener una masa muscular y un tamaño de aleta más pronunciados. En contraste, las especies Adelie y Chinstrap se caracterizan por tener menor desarrollo en estos atributos, pero poseen un pico más profundo. Dentro de estas últimas especies también se pueden apreciar diferencias: los pingüinos Chinstrap tienden a tener un tamaño de pico ligeramente mayor en comparación con los Adelie.

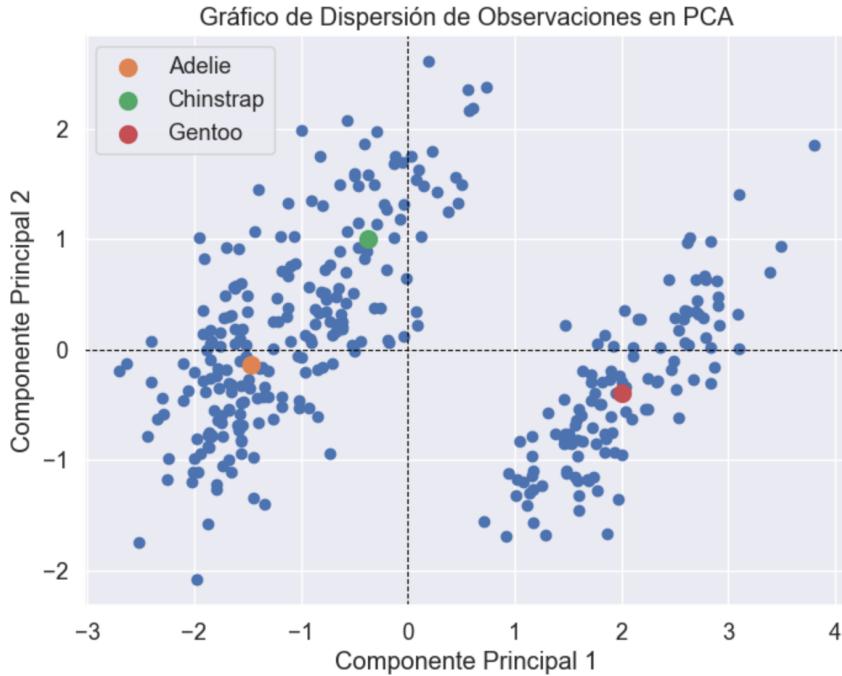


Imagen 9: Dispersión centroides.

Ahora, guiándose de la imagen 4, se podría obtener los índices para los dos componentes, dando como resultado:

$$\text{Componente 1: } (0.455250 \times \text{bill_length_mm_z}) - (0.400335 \times \text{bill_depth_mm_z}) + (0.576013 \times \text{flipper_length_mm_z}) + (0.548350 \times \text{body_mass_g_z})$$

$$\text{Componente 2: } (0.597031 \times \text{bill_length_mm_z}) + (0.797767 \times \text{bill_depth_mm_z}) + (0.002282 \times \text{flipper_length_mm_z}) + (0.084363 \times \text{body_mass_g_z})$$

Si se desea realizar lo mismo por cada especie de pingüino quedaría de la siguiente forma:

	Autovector 1	Autovector 2
bill_length_mm_z	0.509083	-0.822682
bill_depth_mm_z	0.450904	0.183711
flipper_length_mm_z	0.412874	0.025157
body_mass_g_z	0.605850	0.537412

Imagen 10: Autovectores 2 componentes para la especie Gentoo

En la imagen 10 se puede observar los autovectores dados para la especie Gentoo, donde la formula quedaría de la siguiente forma:

Componente 1: $(0.509083 \times \text{bill_length_mm_z}) + (0.450904 \times \text{bill_depth_mm_z}) + (0.412874 \times \text{flipper_length_mm_z}) + (0.605850 \times \text{body_mass_g_z})$

Componente 2: $-(0.822682 \times \text{bill_length_mm_z}) + (0.183711 \times \text{bill_depth_mm_z}) + (0.412874 \times \text{flipper_length_mm_z}) + (0.605850 \times \text{body_mass_g_z})$

	Autovector 1	Autovector 2
bill_length_mm_z	-0.402940	0.378165
bill_depth_mm_z	-0.610229	-0.739973
flipper_length_mm_z	-0.324687	0.509665
body_mass_g_z	-0.599865	0.222874

Imagen 11: Autovectores 2 componentes para la especie Adelie

En la imagen 11 se puede observar los autovectores dados para la especie Adelie, donde la formula quedaría de la siguiente forma:

Componente 1: $-(0.402940 \times \text{bill_length_mm_z}) - (0.610229 \times \text{bill_depth_mm_z}) - (0.324687 \times \text{flipper_length_mm_z}) - (0.599865 \times \text{body_mass_g_z})$

Componente 2: $(0.378165 \times \text{bill_length_mm_z}) - (0.739973 \times \text{bill_depth_mm_z}) + (0.509665 \times \text{flipper_length_mm_z}) + (0.222874 \times \text{body_mass_g_z})$

	Autovector 1	Autovector 2
bill_length_mm_z	0.570924	0.696853
bill_depth_mm_z	0.559406	0.047227
flipper_length_mm_z	0.432181	-0.589067
body_mass_g_z	0.417529	-0.406405

Imagen 12: Autovectores 2 componentes para la especie Chinstrap

En la imagen 12 se puede observar los autovectores dados para la especie Chinstrap, donde la formula quedaría de la siguiente forma:

Componente 1: $(0.570924 \times \text{bill_length_mm_z}) + (0.559406 \times \text{bill_depth_mm_z}) + (0.432181 \times \text{flipper_length_mm_z}) + (0.417529 \times \text{body_mass_g_z})$

Componente 2: $(0.696853 \times \text{bill_length_mm_z}) + (0.047227 \times \text{bill_depth_mm_z}) - (0.589067 \times \text{flipper_length_mm_z}) - (0.406405 \times \text{body_mass_g_z})$

4. Agrupamiento jerárquico

Es crucial destacar que, en el proceso de identificación del agrupamiento jerárquico, se emplearon dos métodos de cálculo de distancia fundamentales: la distancia euclídea y el método de Ward. Estas técnicas son esenciales para determinar la proximidad y la similitud entre los puntos de datos, lo cual es clave para formar clusters coherentes y significativos.

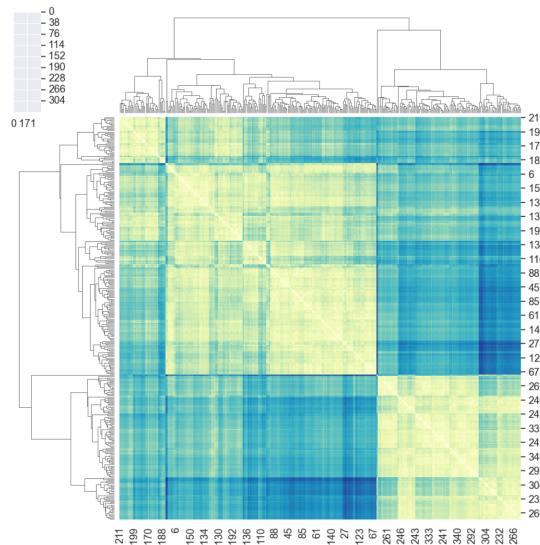


Imagen 13: Mapa de calor del clustering

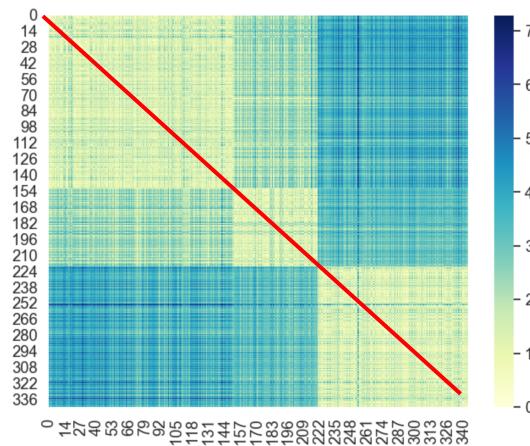


Imagen 14: Matriz de distancias visual entre observaciones

En la imagen 13, se presenta un mapa de calor que visualiza de manera efectiva la magnitud de los valores. Además, en los ejes de este mapa, se observa las llaves que sugieren una agrupación de los atributos. Esta disposición proporciona una percepción de cómo los atributos podrían estar agrupados.

Por otro lado, al realizar un corte imaginario a lo largo de la diagonal del gráfico mostrado en la imagen 14, se revela a través de cambios en los colores, la potencial cantidad de clusters. Este cambio de colores a lo largo de la diagonal actúa como un indicador visual de cómo los datos podrían estar naturalmente divididos en distintos grupos.

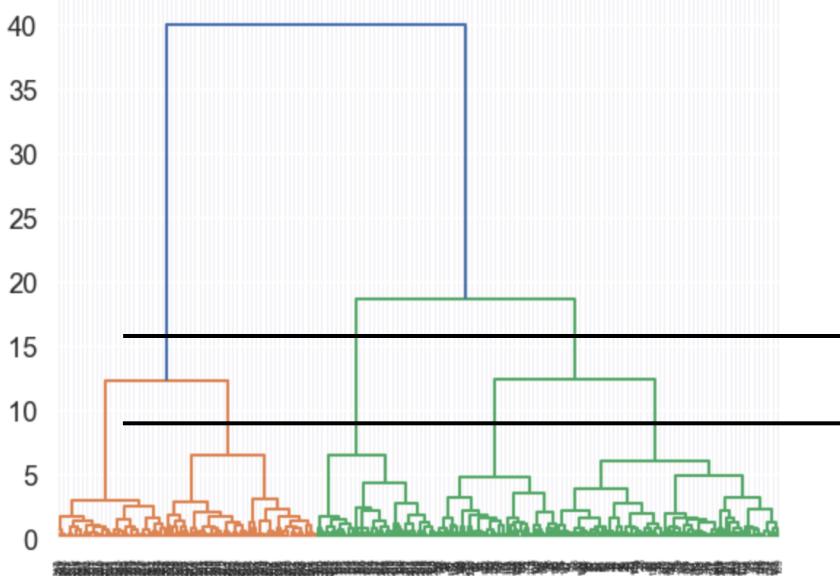


Imagen 15: Dendograma con dos posibles cortes de altura

La imagen 15 ofrece una representación gráfica más detallada y enfocada que la imagen 13, centrándose específicamente en las llaves. En este grafico se puede observar la altura, en la cual se han propuesto dos líneas de corte potenciales. El primer corte sugiere la formación de tres clusters, mientras que el segundo indica la posibilidad de obtener cinco clusters.

5. Agrupamiento K-Means

El agrupamiento mediante el algoritmo K-Means se llevó a cabo experimentando con diferentes cantidades de clusters, realizando un total de 11 iteraciones para evaluar los resultados variados como se puede ver en la imagen 16.

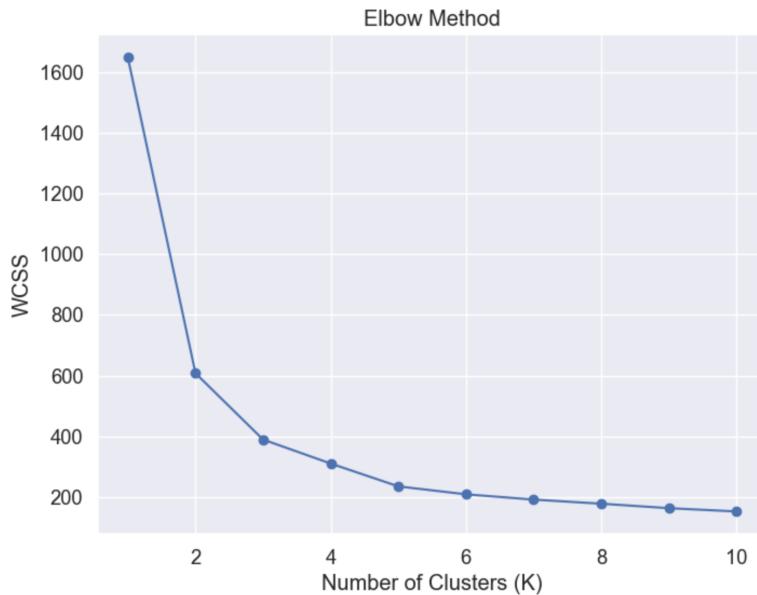


Imagen 16: Gráfica del método del Codo

A través de este proceso, se observó que la formación del "codo", indicativo del número óptimo de clusters, se empezaba a manifestar entre el tercer y cuarto cluster. Sin embargo, cabe destacar que la disminución de la suma de los cuadrados dentro de los clusters (WCSS) comenzó a ser marginal después del quinto cluster. Este patrón sugiere que, aunque el punto de inflexión se encuentra entre el tercer y cuarto cluster, se podría extender el número de clusters hasta el quinto, pero no más allá, ya que no conlleva mejoras significativas en términos de la compactación de los clusters.

6. Validación del Agrupamiento



Imagen 17: Método de la silueta

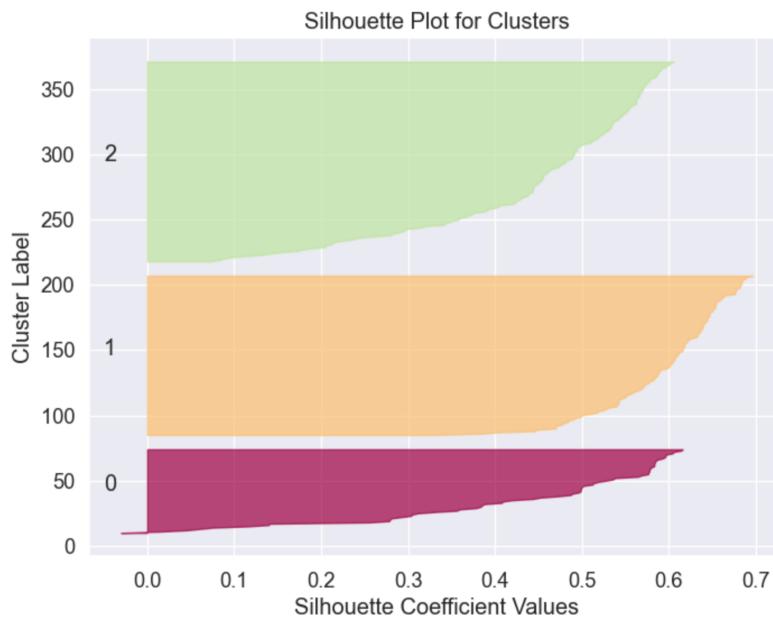


Imagen 18: Método de la silueta por cluster

La imagen 18 presenta un análisis detallado de los resultados del agrupamiento con los tres clusters. Al examinar la gráfica de silueta correspondiente, se destaca que el rango óptimo de agrupamiento se encuentra entre en el grupo 2. Esto se determina basándose en el coeficiente de silueta, donde valores más cercanos a 1 indican una mejor calidad de agrupamiento. En este rango específico, los coeficientes de silueta alcanzan sus niveles más altos, lo que sugiere que los grupos formados son cohesivos internamente y bien separados entre sí, proporcionando así una estructuración clara y efectiva del conjunto de datos.

7. Comparación de los resultados (Jerárquico vs K-Means)

Como se demostró en los puntos 4, 5 y 6, ambos métodos de agrupamiento, jerárquico y K-Means, resultaron eficaces para lograr una segmentación adecuada en clusters. La elección del número de clusters a utilizar depende en gran medida del criterio del analista de datos o de los requerimientos específicos del análisis. En el contexto de esta investigación, el método K-Means resultó ser más preferible, gracias a su enfoque intuitivo y a las herramientas disponibles que facilitan la validación de las decisiones tomadas.

En términos de los resultados obtenidos, ambos métodos mostraron similitudes notables, aunque con una diferencia destacable. El método jerárquico tendió a favorecer una división en 5 clusters, mientras que el enfoque K-Means se inclinó hacia 3 o 4 clusters. Esta diferencia puede reflejar cómo cada método procesa y interpreta los datos, destacando la importancia de seleccionar el método más adecuado en función de los objetivos y peculiaridades del conjunto de datos específico en estudio.

8. Proporciona una interpretación de los grupos.

La imagen 19 ilustra de manera notable cómo los clusters pueden interpretarse como representaciones de las diferentes especies de pingüinos. Esta correlación se debe a que las características estudiadas están fuertemente asociadas con las especies específicas. Como se observó anteriormente, es factible realizar una agrupación que exceda el número de especies existentes, optando por 4 o 5 clusters, y aun así obtener resultados estadísticamente robustos.

Este fenómeno se puede explicar considerando que las especies de pingüinos Adelie y Chinstrap comparten características tan similares que sugieren la posibilidad de un ancestro común. Estas especies podrían estar experimentando una divergencia evolutiva continua, desarrollando rasgos cada vez más distintivos. Debido a su proximidad en términos de características, podrían ser agrupadas en uno o dos clusters adicionales, aparte de los centroides que representan a cada especie individualmente. Esta interpretación de los datos apunta a una relación evolutiva interesante y sugiere un patrón de segregación progresiva entre estas especies de pingüinos.

La posibilidad de subdividir los datos en 5 clusters también sugiere un enfoque interesante en el análisis ya que la especie Gentoo podría ser segregada. Como se evidencia en la imagen 19, esta especie presenta una variabilidad significativa en sus características, lo cual justifica la consideración de una segregación adicional. La diversidad observada en los rasgos de los pingüinos Gentoo indica que podrían ser divididos en subgrupos más específicos. Esta subdivisión adicional podría proporcionar una comprensión más detallada y refinada de las diferencias intrínsecas dentro de la especie, revelando patrones o subespecies que podrían no ser evidentes en un análisis menos granular.

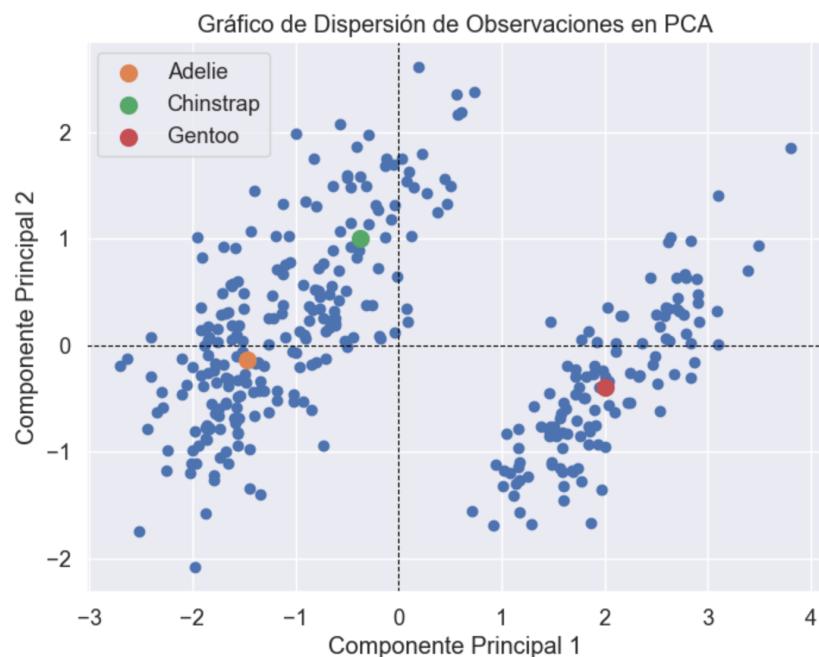


Imagen 19: Gráfico de dispersión con centroide de especies

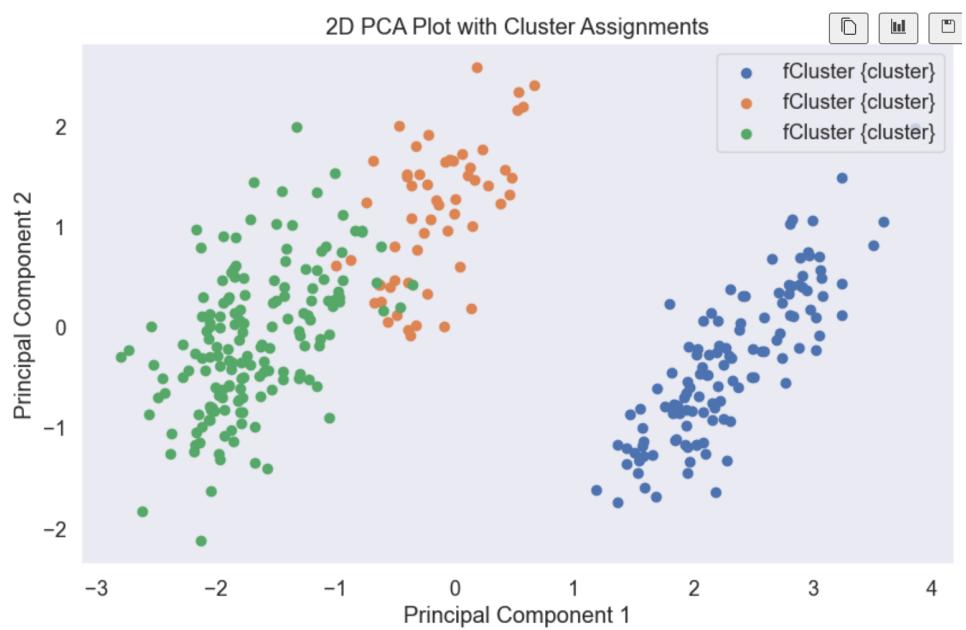


Imagen 20: Gráfico de dispersión con 3 clusters

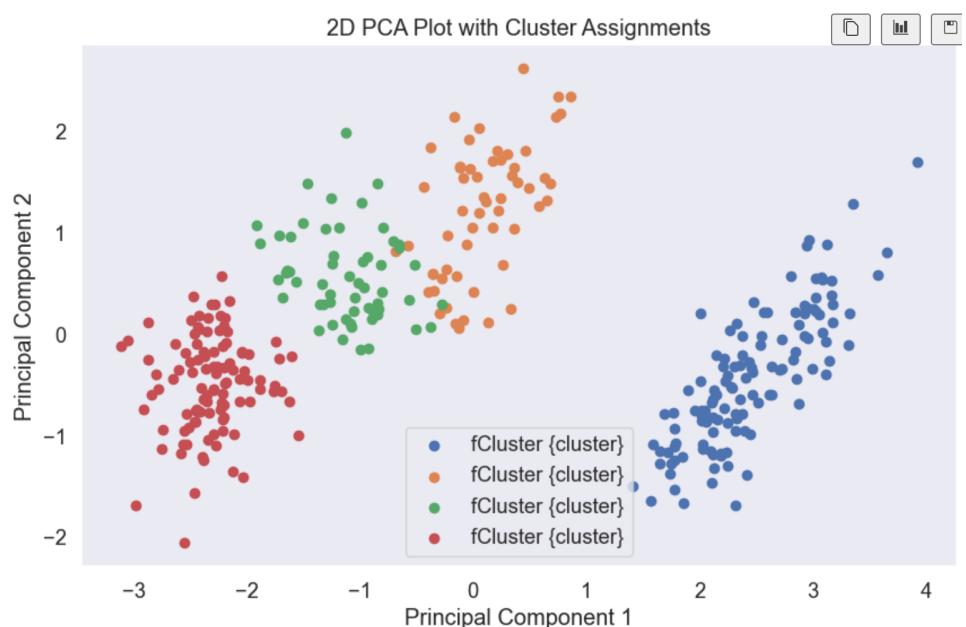


Imagen 21: Gráfico de dispersión con 4 clusters

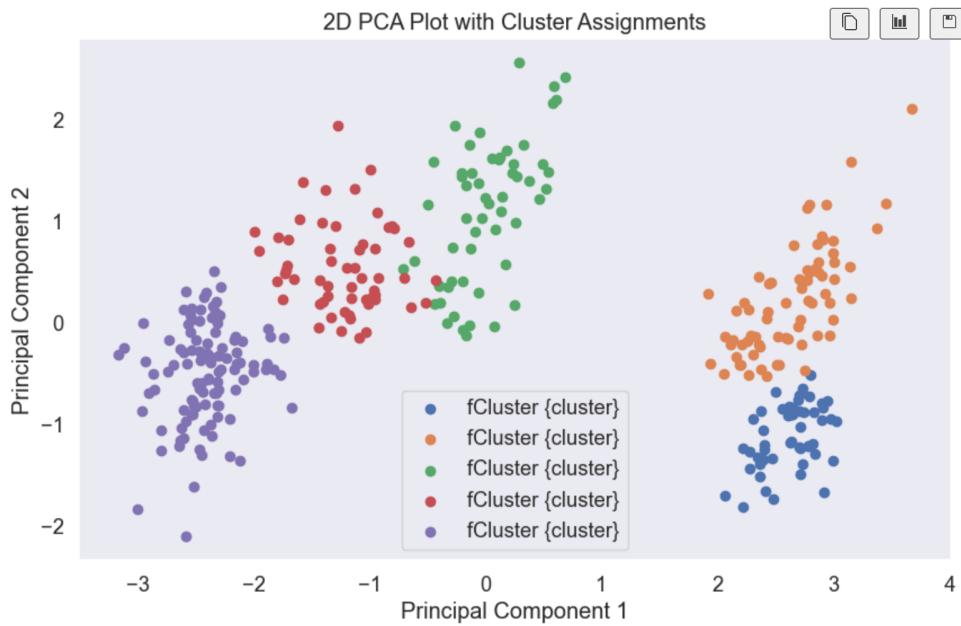


Imagen 22: Gráfico de dispersión con 5 clusters

9. Conclusión, limitaciones y desafíos

En conclusión, el análisis realizado ha revelado patrones significativos y relaciones dentro del conjunto de datos penguins como sería:

- La aplicación de Análisis de Componentes Principales (ACP) ha permitido una reducción efectiva de la dimensionalidad, facilitando una visualización más clara y la identificación de agrupaciones naturales en los datos.
- De los datos se puede notar una clara correlación entre el tamaño del pico con el tamaño de la aleta y la masa muscular, mientras que la profundidad del pico viene inversamente relacionada con los atributos anteriormente mencionados.
- Las técnicas de clustering aplicadas han demostrado ser valiosas en la categorización de las especies de pingüinos, destacando diferencias y similitudes clave en sus características físicas.
- Las técnicas de clusterización por jerarquía y K-means pueden llegar a un punto muy parecido, por lo que la aplicación de cualquiera de las dos supone un buen acercamiento, pero con una leve inclinación al K-means por el hecho de tener métodos de verificación.
- la interpretación de los clusters puede ser subjetiva y dependiente de la elección del número de clusters y del algoritmo de clustering utilizado
- Este estudio subraya la importancia de la ciencia de datos en el análisis en el ámbito biológico y abre puertas a futuras investigaciones que podrían explorar más a fondo las dinámicas de vida de estos fascinantes animales.
- Como limitaciones se puede resaltar la cantidad de especies de pingüinos que posee el dataset, así como también la dimensionalidad de los datos y la integridad de los mismos.