

Interpretación de Resultados

El análisis de los datos del archivo de los pingüinos permitió identificar diferencias físicas importantes entre los individuos, especialmente en variables como la longitud y profundidad del pico, la longitud de la aleta y la masa corporal. Estas características se usaron para aplicar el algoritmo K-means. Antes de elegir el número de grupos, se utilizó el método del codo, donde se observó que la inercia disminuye notablemente hasta $k = 3$, y a partir de ese punto la curva se vuelve casi plana. Esto indicó que $k = 3$ era el valor más adecuado, ya que agregar más clusters no aportaría una separación significativamente mejor. Por lo tanto, la elección de k no fue arbitraria, sino justificada por la gráfica.

Al observar los centros de los clusters, se identificaron tres perfiles físicos claros: un grupo de pingüinos más grandes con picos largos, aletas largas y mayor peso; otro grupo conformado por pingüinos más pequeños con menor masa corporal, y un tercer grupo intermedio. Estos centros pueden considerarse representativos, ya que reflejan valores promedio coherentes con los datos reales y muestran diferencias visibles entre los grupos. Si se hubiera usado un valor de k menor, como 1 o 2, se habría perdido esta diferenciación, mezclando pingüinos distintos en un solo grupo. Por otro lado, aumentar demasiado el valor de k dividiría los grupos en subgrupos demasiado similares entre sí, haciendo que los centros dejaran de tener un significado claro y útil.

En cuanto a la distancia entre los centros, se observó que no están excesivamente cercanos, lo que indica que los clusters son distintos entre sí y no se traslapan demasiado. Esto respalda que la segmentación fue adecuada. Sin embargo, es importante considerar que si en los datos hubiera muchos valores atípicos (outliers), los centros podrían haberse desplazado hacia esos valores extremos, ya que K-means calcula promedios. En ese caso, los clusters podrían perder sentido, lo que demuestra la importancia de revisar y limpiar los datos antes de aplicar el algoritmo.

La gráfica de dispersión refuerza la interpretación: aunque no se aplicaron colores para distinguir los clusters, se puede observar que los datos se agrupan de forma visible en tres zonas, lo cual coincide con los tres grupos encontrados. Esto confirma que la agrupación no fue producida al azar, sino que las variables realmente explican diferencias estructurales entre los pingüinos.

En resumen, el análisis muestra que la población estudiada no es homogénea, sino que presenta tres tipos corporales bien definidos. La elección de $k = 3$ fue adecuada, los centros son representativos, los grupos están claramente diferenciados y el modelo permite describir variaciones reales dentro de la especie.

