PCA y Clusterización de poblaciones de pingüinos

Minería de datos y Modelización Predictiva

Bartolomé Mestre Fons

Miércoles, 16 de Enero de 2025

Minería de datos y Modelización Predictiva
2024-2025
UCM - Universidad Complutense de Madrid

Índice

1. Introducción

En el siguiente documento se pretende analizar una serie de poblaciones de pingüinos para diferentes especies, sexos e islas donde se tomó la muestra. Se realizará primeramente un Análisis por Componentes Principales (PCA) sobre los datos con el fin de identificar similitudes/discrepancias entre las características fisionómicas de las poblaciones. A continuación, reutilizando los resultados obtenidos en el PCA, se propone la agrupación de los datos en clústers mediante métodos Jerárquicos y No Jerárquicos, utilizando diferentes métricas a la hora de escoger el número de clústers óptimo. Finalmente, se evalua cada agrupación y se caracteriza cada clúster basándose en estadísticos descriptivos y sus centroides. Se adjunta la bibliografía usada al final del documento.

2. PCA

2.1. Matriz de correlación

Esta sección corresponde al apartado 1 del PCA de la hoja de ejercicios.

Los datos a usar en el presente documento tienen la siguiente apariencia:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Adelie	Torgersen	39.1	18.7	181	3750	Male
Adelie	Torgersen	39.5	17.4	186	3800	Female
Adelie	Torgersen	40.3	18.0	195	3250	Female
Adelie	Torgersen	36.7	19.3	193	3450	Female
Adelie	Torgersen	39.3	20.6	190	3650	Male

Cuadro 1: Muestra de las 5 primeras observaciones.

Se tiene un total de 333 observaciones sin valores nulos detectados, donde cada observación vendrá descrita por un total de 7 características:

1. **species**: indica la especie del pingüino. Puede ser Adelie, Chinstrap o Gentoo.
2. **island**: indica la isla donde se tomó la observación. Puede ser Torgersen, Biscoe o Dream. Se tomaron medidas de especímenes Adelie en las 3 islas, mientras que las observaciones de las especies Chinstrap y Gentoo se limitan a las islas Dream y Biscoe, respectivamente.
3. **bill_length_mm**: indica indica la longitud del pico en milímetros.
4. **bill_depth_mm**: indica indica la profundidad del pico en milímetros.
5. **flipper_length_mm**: indica indica la longitud de la aleta en milímetros.
6. **body_mass_g**: indica indica la masa corporal del pingüino en gramos.
7. **sex**: describe el sexo del pingüino, que puede ser Male (macho) o Female (hembra).

Se calcula a continuación la matriz de correlación entre las variables numéricas mencionadas arriba con el fin de identificar posibles efectos de colinealidad:

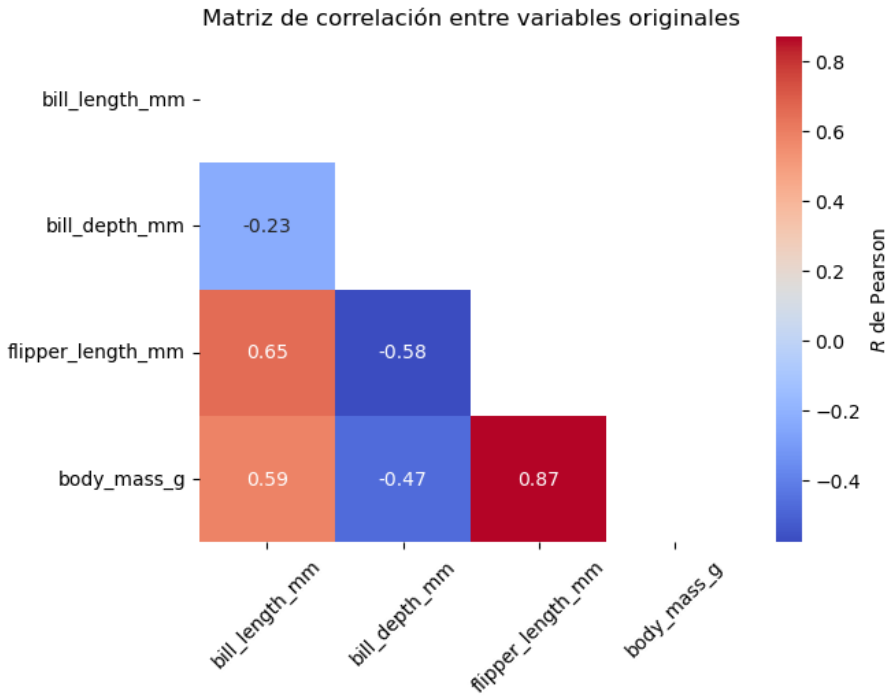

Figura 1: Matriz de correlación entre las variables numéricas originales. Se elimina el triángulo superior para reducir la redundancia de valores.

Se observa en la matriz que aquellas parejas de variables con mayor correlación positiva son body_mass_g/ flipper_length_mm, flipper_length_mm/bill_length_mm y body_mass_g/bill_length_mm (0.87, 0.65 y 0.59, respectivamente). Por otro lado, se identifica la pareja flipper_length_mm/bill_depth_mm como la que posee una mayor correlación negativa (-0.58). Así, uno puede inferir que aquellos individuos con un peso corporal elevado poseerán aletas con longitudes por encima de la media total, a la par que especies con picos poco profundos tendrán aletas grandes. Una vez observadas las dependencias entre las variables, estamos en posición de realizar un PCA sobre los datos.

2.2. Estudio de las Componentes Principales

Esta sección corresponde al apartado 2 del PCA de la hoja de ejercicios.

En primer lugar, se estandarizan las variables numéricas con el fin de fijar una misma escala, tal que para cada variable la media sea 0 y la desviación estándar 1. Se muestra a continuación el resultado:

bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	species	island	sex
-0.90	0.78	-1.43	-0.57	Adelie	Torgersen	Male
-0.82	0.12	-1.07	-0.51	Adelie	Torgersen	Female
-0.67	0.42	-0.43	-1.19	Adelie	Torgersen	Female
-1.34	1.09	-0.57	-0.94	Adelie	Torgersen	Female
-0.86	1.75	-0.78	-0.69	Adelie	Torgersen	Male

Cuadro 2: Muestra de las 5 primeras observaciones estandarizadas.

A continuación, se procede a realizar el PCA con las herramientas que dispone la librearía sklearn. Se toman para empezar 4 Componentes Principales (tantas como el número de variables originales). Así, los autovalores resultantes son los siguientes:

	Componente 1	**Componente 2**	**Componente 3**	**Componente 4**
Autovalor	2.75	0.78	0.37	0.11

Cuadro 3: Autovalores de las Componentes Principales.

Uno observa como el autovalor asociado a la primera Componente Principal es el más grande, por lo que dicha componente será la que explique la mayor parte de la variabilidad total de las variables originales. En este punto, se debe graficar la Variabilidad Explicada por Componente Principal con el fin de usar el *método del codo* para escoger el número de Componentes Principales adecuado:

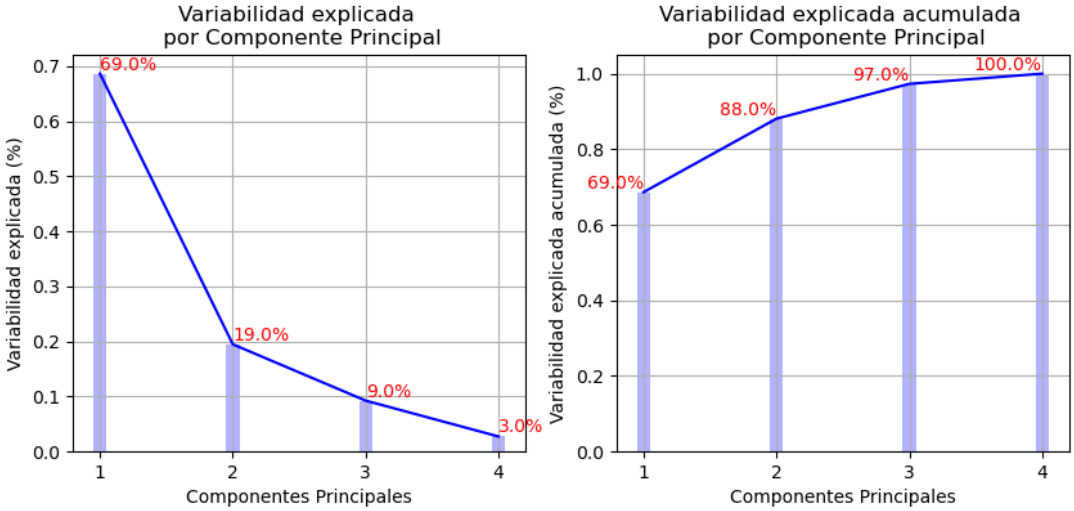

Figura 2: Variabilidad Explicada y Variabilidad Explicada Acumulada en función del número de Componentes Principales escogido.

Se observa la existencia de un cambio de tendencia en el gráfico de la Variabilidad Explicada allí donde se consideran las 2 Componentes Principales, siendo este el número adecuado a escoger para el futuro análisis. Así, se consigue explicar el 88 % de la variabilidad total (69 % y 19 % por la Componente Principal 1 y 2, respectivamente). Así, se desplaza el análisis de 4 variables originales a tan solo 2 Componentes Principales, reduciendo la complejidad de los datos y posibilitando una representación gráfica de los mismos. Cabe aquí mencionar que, como se verá más adelante, es aconsejable tomar (en contra del criterio del *método del codo*) 3 Componentes Principales para indagar con mayor profundidad en las relaciones/discrepancias entre las observaciones con mismo/igual especie, sexo o isla. Así, el futuro análisis se mantendrá para 2 Componentes Principales, y se comentará superficialmente el caso para 3 Componentes Principales.

2.3. PCA con 2 Componentes Principales

2.3.1. Cálculo manual de las Componentes Principales

Esta sección corresponde al apartado 3.a del PCA de la hoja de ejercicios.

Se detalla en esta subsección el desarrollo matemático para el cálculo de la primera Componente Principal de la primera observación, para así poder aplicarlo al resto. Una vez determinado el número de Componentes Principales a considerar, se hallan los autovalores y autovectores de la matriz de correlación (hallada previamente con la ecuación $R = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_i)(X_i - \bar{X}_i)^T$, donde n es el número de variables originales). A continuación, se hallan los autovalores resolviendo la siguiente ecuación matricial:

$$(R - \lambda I) = 0 \rightarrow \begin{pmatrix} 1 & -0{,}23 & 0{,}65 & 0{,}59 \\ -0{,}23 & 1 & -0{,}58 & -0{,}47 \\ 0{,}65 & -0{,}58 & 1 & 0{,}87 \\ 0{,}59 & -0{,}47 & 0{,}87 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{1}$$

Los autovalores son los hallados en las dos primeras columnas de 3, donde cada autovalor λ_i explica un $\frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j} \cdot 100$ por ciento de la variabilidad total. Se procede a continuación a calcular los autovectores asociados, resolviendo la siguiente ecuación:

$$(R - \lambda_i I)x = 0 \tag{2}$$

Aquí, λ_i es cada autovalor según cuantas Componentes Principales se halla decidido mantener. Para el caso de solo 2, se deben resolver las siguientes 2 ecuaciones matriciales:

$$\left[\begin{pmatrix} 1 & -0{,}23 & 0{,}65 & 0{,}59 \\ -0{,}23 & 1 & -0{,}58 & -0{,}47 \\ 0{,}65 & -0{,}58 & 1 & 0{,}87 \\ 0{,}59 & -0{,}47 & 0{,}87 & 1 \end{pmatrix} - 2{,}75 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\right] \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ t_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3}$$

$$\left[\begin{pmatrix} 1 & -0{,}23 & 0{,}65 & 0{,}59 \\ -0{,}23 & 1 & -0{,}58 & -0{,}47 \\ 0{,}65 & -0{,}58 & 1 & 0{,}87 \\ 0{,}59 & -0{,}47 & 0{,}87 & 1 \end{pmatrix} - 0{,}78 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\right] \begin{pmatrix} x_2 \\ y_2 \\ z_2 \\ t_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{4}$$

$W_1 = (x_1, y_1, z_1, t_1)$ y $W_2 = (x_2, y_2, z_2, t_2)$ son los autovectores de las Componentes Principales 1 y 2, respectivamente. Los autovectores hallados son los mostrados a continuación:

	Autovector 1	**Autovector 2**
bill_length_mm_z	0.45	0.60
bill_depth_mm_z	-0.40	0.80
flipper_length_mm_z	0.58	0.01
body_mass_g_z	0.55	0.08

Cuadro 4: Tabla de Autovectores de las dos primeras Componentes Principales.

Con esto, ya se puede calcular la primera Componente Principal para la primera observación $CP_{1,1}$, siguiendo la siguiente ecuación:

$$CP_{i,j} = Z_j W_i \tag{5}$$

Aquí, Z_j es un vector fila conteniendo el valor de las variables originales estandarizadas de la observación j y W_i es el autovector de la Componente Principal i. Para la primera observación, se tiene lo siguiente:

$$CP_{1,1} = Z_1 W_1 = Z_1^{(1)} W_1^{(1)} + Z_1^{(2)} W_1^{(2)} + Z_1^{(3)} W_1^{(3)} + Z_1^{(4)} W_1^{(4)} = \tag{6}$$

$$(-0{,}90) \cdot 0{,}45 + 0{,}78 \cdot (-0{,}40) + (-1{,}43) \cdot 0{,}58 + (-0{,}57) \cdot 0{,}55 = -1{,}85 \tag{7}$$

2.3.2. Estudio de las Variables Originales en función de las Componentes Principales

Esta sección corresponde al apartado 3.b y 3.d del PCA de la hoja de ejercicios.

Se analiza en la presente subsección como deben interpretarse las Variables Originales en función de las Componentes Principales. Para ello, se grafica en un mapa de calor la correlación entre las variables originales y las nuevas componentes halladas:

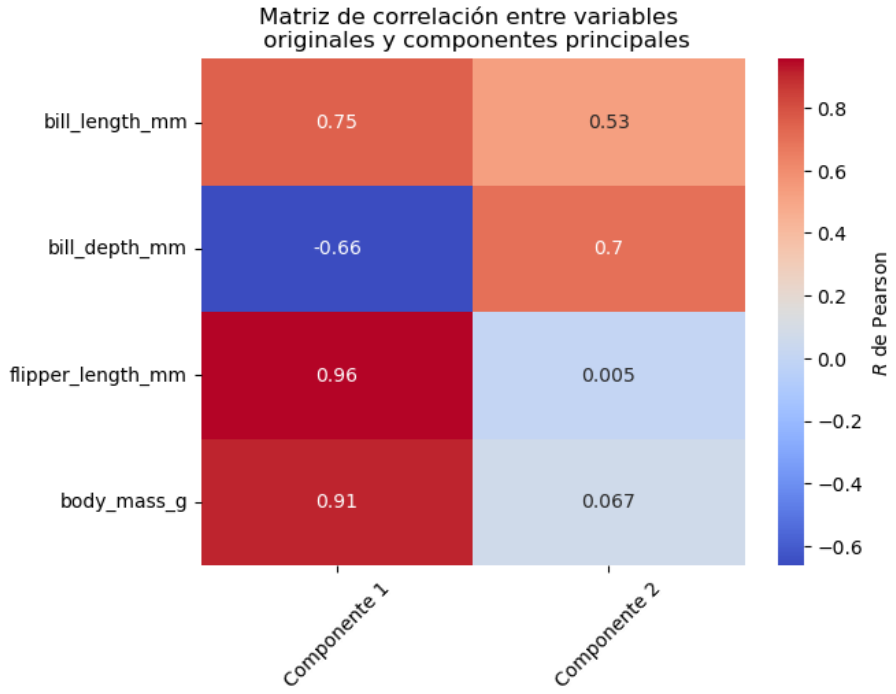

Figura 3: Matriz de Correlación entre las variables originales y las CPs.

Se observan fuertes correlaciones tanto positivas como negativas entre la primera Componente y las variables originales, lo que conduce a que la primera Componente Principal influencia fuertemente el comportamiento de las variables originales (con coeficientes de Pearson que abarcan un rango de 0.96 a -0.66). Esta describe sobre todo las variables body_mass_g y flipper_length_mm (R de Pearson de 0.91 y 0.96, respectivamente), por lo que se concluye que estas dos variables están representadas de manera directa por esta componente (y viceversa, la primera CP viene descrita en mayor parte por estas dos variables originales). Por otro lado, la segunda Componente Principal apenas describe las dos primeras variables pero si las dos otras medianamente (0.53 para bill_length_mm y 0.7 bill_depth_mm), lo que indica que estas dos últimas se ven influenciadas tanto por la primera CP como la segunda CP en grados similares.
Si se elevan estas correlaciones al cuadrado, se obtienen los conocidos *cosenos al cuadrado*, que representan la proporción de variabilidad que cada componente explica de cada variable. Se grafica en un mapa de calor tal y como sigue:

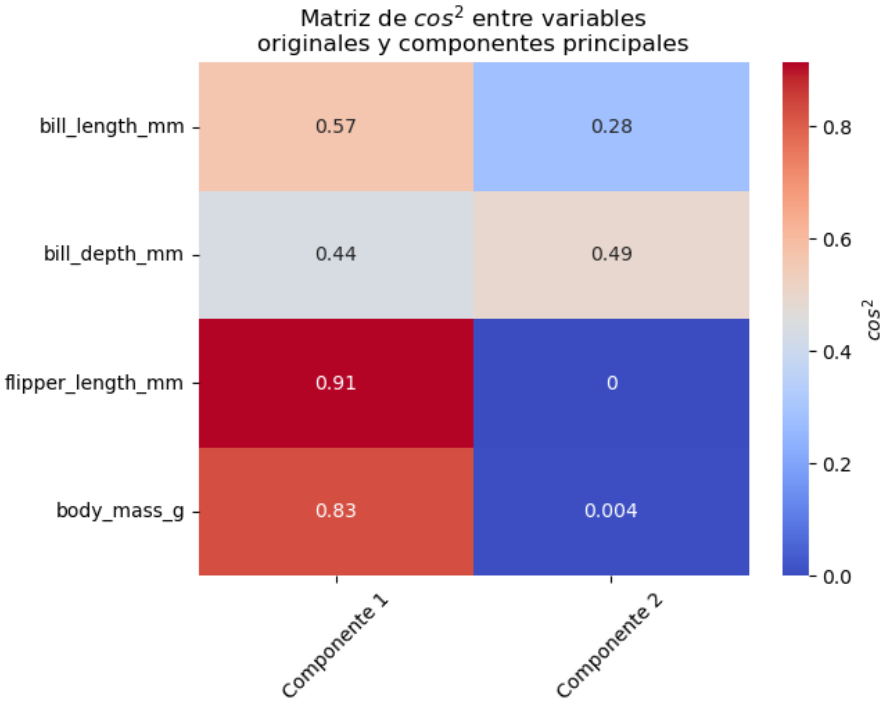

Figura 4: *Cosenos al cuadrado* de las variables originales en función de las Componentes Principales.

En efecto, la primera Componente describe directamente las variables body_mass_g y flipper_length_mm, mientras que para las otras dos variables las dos Componeentes Principales las describen de manera similar. Esto se discierne de manera más visual con un gráfico vectorizado de las variables sobre el plano de las Componentes Principales:

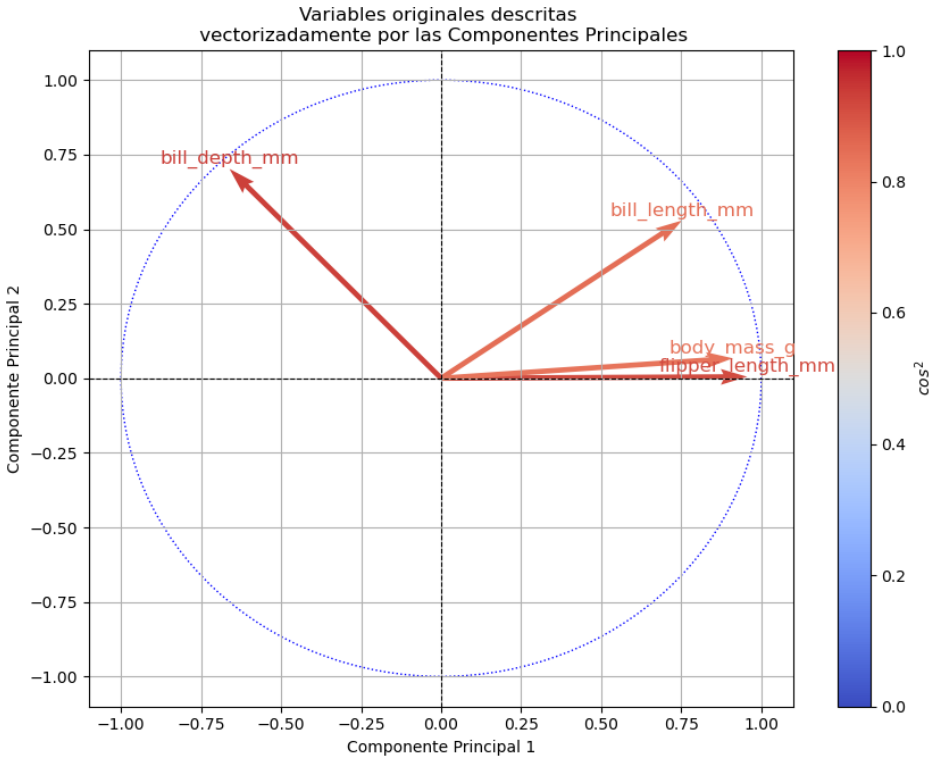

Figura 5: Variables originales vectorizadas en función de las Componentes Principales.

Se representa un vector por cada variable original sobre el plano de las Componentes Principales, donde la longitud del vector viene dada por la suma de todos los cos^2 de una misma variable para todas las CPs, lo que determina cuán bien describen las CPs la variabilidad de la variable. Se observa como las variables body_mass_g y flipper_length_mm se hallan casi horizontales apuntando hacia valores positivos de la primera CP. Este fenómeno apoya el hecho de que la primera componente describa casi enteramente el comportamiento de estas dos variables. Por otra parte, se vuelve a confirmar que el comportamiento de las otras dos variables restantes (bill_depth_mm y bill_length_mm) viene descrito de manera similar por ambas componentes, ambas con similar contribución positiva por la segunda CP pero con contribuciones de diferente signo en cuanto a la primera CP (pero en misma proporción). El diagrama también debe leerse en el sentido inverso, esto es, la Componente Principal 1 viene fuertemente influenciada por body_mass_g y flipper_length_mm y medianamente por bill_depth_mm y bill_length_mm. Para la segunda Componente Principal, su influecia proviene casi enteramente de estas dos últimas variables.
A partir de las proposiciones anteriores, uno puede concluir que aquellos individuos de la población que posean mayores pesos corporales o longitudes de las aletas poseerán un valor elevado de la Componente Principal 1 y a la inversa para pesos o tamaños menores. Para estos casos, la Componente Principal tendrá un valor bajo cercano a 0. Para otros individuos con tamaños y profundidades del pico elevadas se tendrán valores para las Componentes Principales 1 y 2 similares entre si en valor absoluto (al formar las variables bill_depth_mm y bill_length_mm un ángulo cercano a 45º respecto a la horizontal).
Finalmente, uno puede estudiar la variabilidad explicada por las Componentes Principales mostrando en un gráfico de barras la suma total de todos los cos^2 por variable:

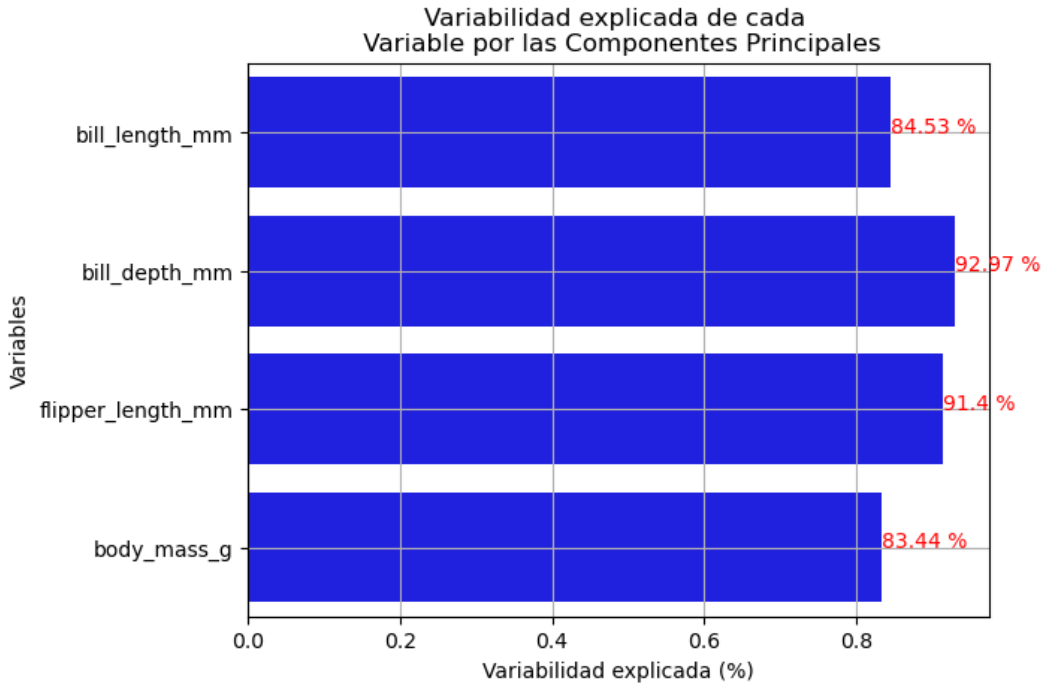

Figura 6: Variabilidad explicada de las Variables originales por las Componentes Principales.

Se observa como las CPs describen bien el comportamiento de las variables bill_depth_ y flipper_length_mm con un 92.97 % y 91.4 % de variabilidad explicada, respectivamente. Para la variable body_mass_g se tiene el peor resultado con un 83.44 %, si bien esto no debe desesperanzar al lector ya que estos resultados son buenos al poder explicar más del 80 % de la variabilidad de las variables (con 3 CPs se habría obtenido una mayor variabilidad explicada a costa de añadir una nueva CP, y con 4 se explica el 100 %).

2.4. Análisis de la población en CPs

Esta sección corresponde al apartado 3.c del PCA de la hoja de ejercicios.

Finalmente, se procede a graficar en los ejes de las PCs las observaciones diferenciando entre las 3 especies:

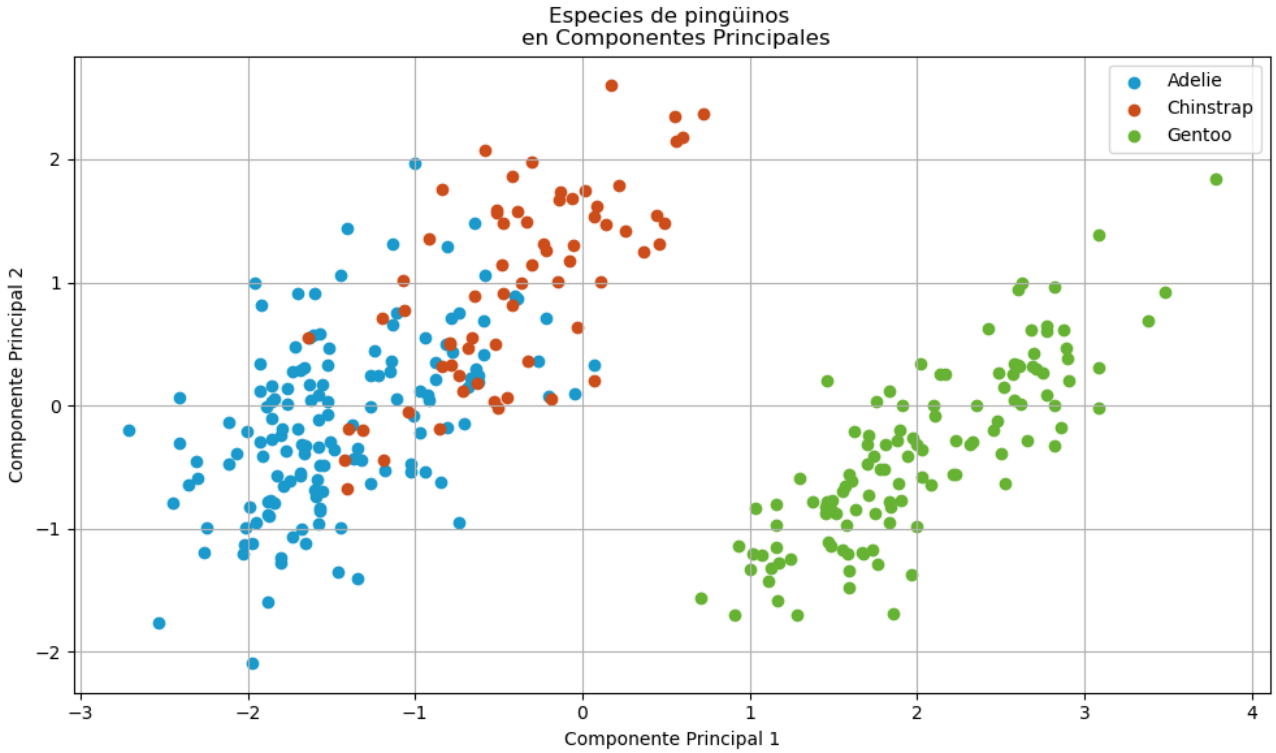

Figura 7: Observaciones en función de las CPs diferenciadas por especie.

Se identifican 2 agrupaciones claramente separadas; a la derecha los individuos de la especie Gentoo y a la izquierda solapándose aquellos de las especies Adelie y Chinstrap. Ateniéndose al significado de las CPs en función de las variables originales, se halla que la especie Gentoo es la que posee una mayor masa corporal y una mayor longitud de aletas (mayor valor de la Componente Principal 1), siendo el caso inverso para las especies Chinstrap y Adelie (Adelie presenta una menor masa corporal y longitud de aleta que la especie Chinstrap). Por otra parte, la especie Gentoo es a la vez la especie con mayor longitud de pico (correlación positiva entre bill_length_mm y Componente 1) pero con menor profundidad (correlación negativa entre bill_depth_mm y Componente 1), siendo el caso inverso para las especies Chinstrap y Adelie, donde algunas observaciones poseen valores negativos en la Componente 1. En la Componente 2 los dos grupos (Gentoo, Adelie y Cinstrap) se hallan distribuidos de manera similar por lo que la segunda CP no aporta diferencias notables entre las 3 especies.

A continuación, se estudia la distribución de las 3 especies en función de su sexo:

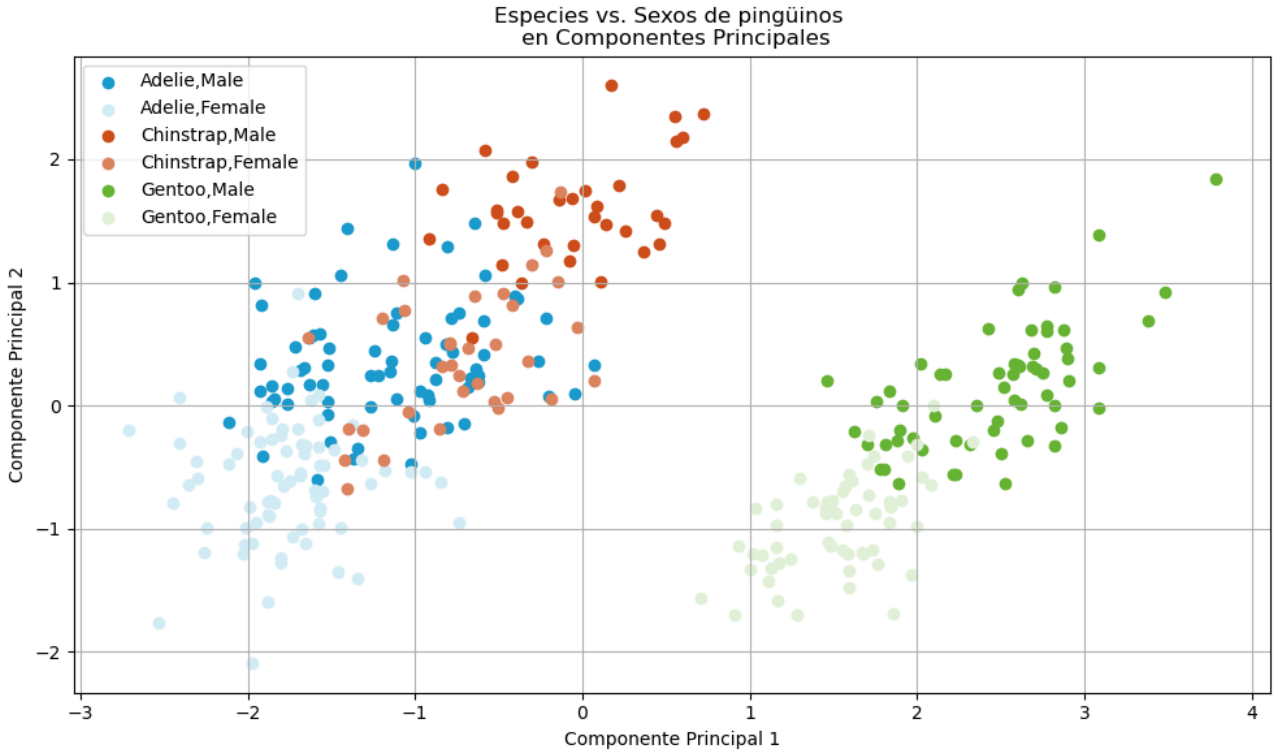

Figura 8: Observaciones en función de las CPs diferenciadas por especie y sexo.

Uno puede observar que al diferenciar entre el sexo del individuo, existen claras evidencias que apuntan a discrepancias fisionómicas entre los individuos machos y hembras. En particular, las hembras presentan menores valores de las Componentes Principales 1 y 2 que sus respectivos machos, por lo que se concluye que las hembras poseen menor peso corporal, menor longitud de las aletas y menor longitud del pico. Ahora bien, recuérdese que la variable que da cuenta de la profundidad del pico se comporta a la inversa de la Componente Principal 1, por lo que se espera que las hembras tengan una mayor profundidad del pico.

Finalmente, se estudian los especímenes en función de la isla donde se tomaron las observaciones:

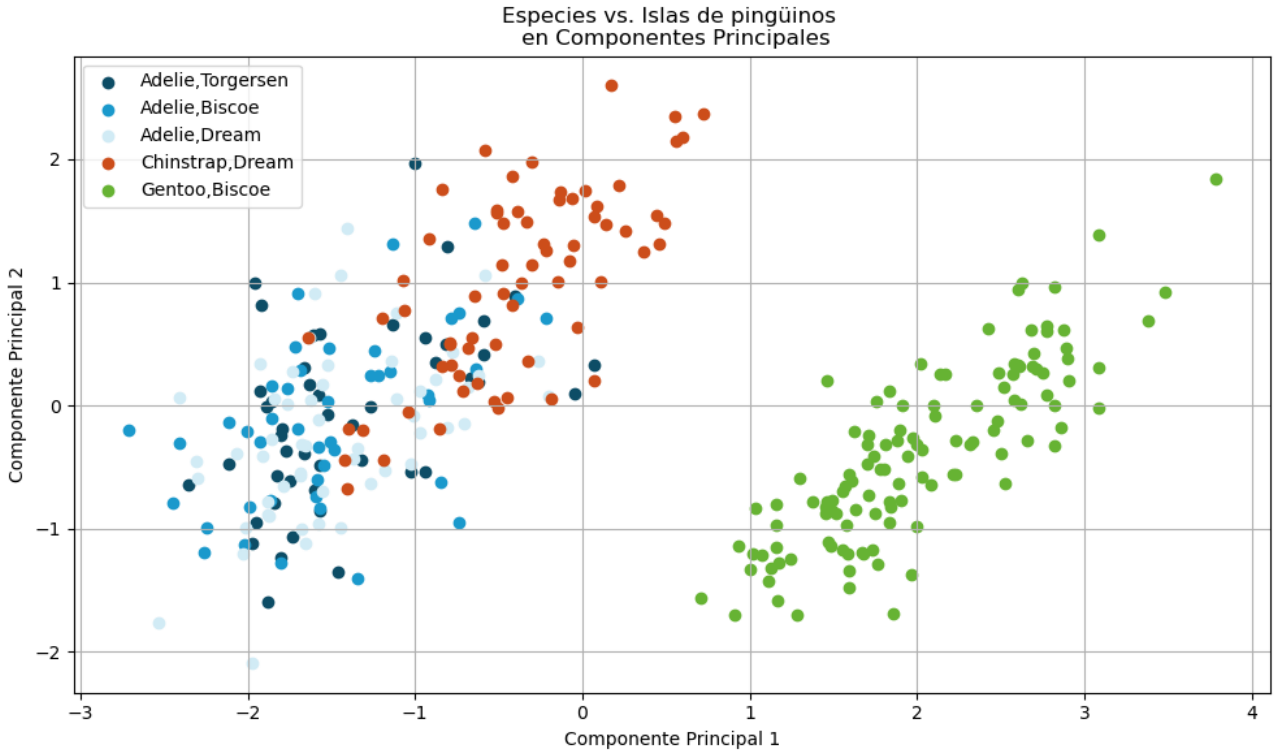

Figura 9: Observaciones en función de las CPs diferenciadas por especie e isla.

Este análisis se ve limitado por la falta de observaciones de las especies Chinstrap y Gentoo en más de una isla, por lo que tan solo se puede analizar la especie Adelie para un total de 3 islas diferentes. Uno observa como las observaciones de la especie Adelie se distribuyen de manera uniforme indiferentemente de la isla de origen, por lo que se concluye que la isla donde la medición se tomó no influye a la hora de las características fisionómicas del espécimen, lo que sugiere que la especie Adelie no es endémica de la región y existe un movimiento de individuos de esta especie entre islas.

Como contenido adicional, se grafica a continuación las observaciones en un mapa 3D en función de 3 Componentes Principales. Si bien la tercera Componente aporta poca variabilidad explicada de las variables originales, permite visualizar las agrupaciones entre especies, sexos e islas de manera más discernible:

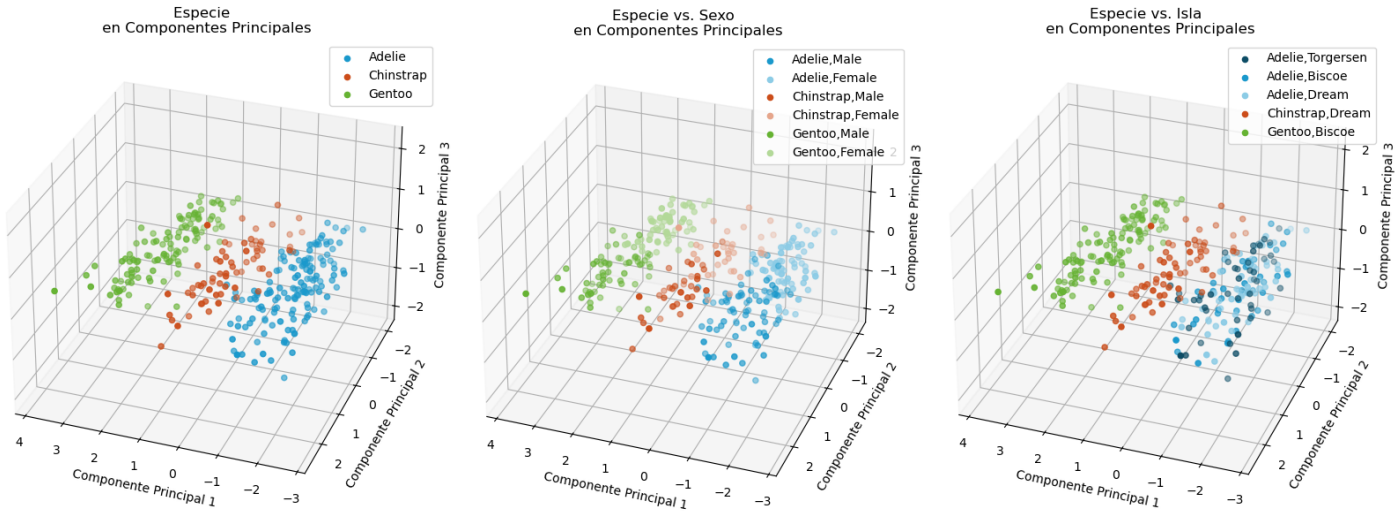

Figura 10: Observaciones en función de las CPs diferenciadas por especie, especie y sexo y especie e isla.

Así, se pueden visualizar discrepancias claras entre la fisionomía de las especies por su sexo, mientras que en el último gráfico se confirma que la isla donde se registró la observación no influye en las características del

individuo.

3. Clusterización

3.1. Matriz de Distancias

Esta sección corresponde al apartado 1 de Clusterización de la hoja de ejercicios.

Se pretende a continuación calcular la matriz de distancias entre las diferentes observaciones. La métrica escogida es la Euclidia ya que las variables son continuas y se hallan en un espacio Euclidio. Dado que se disponen de 333 observaciones, se limita el análisis a las 33 primeras observaciones (una décima parte del total). Se muestra la matriz de distancias en un mapa de calor:

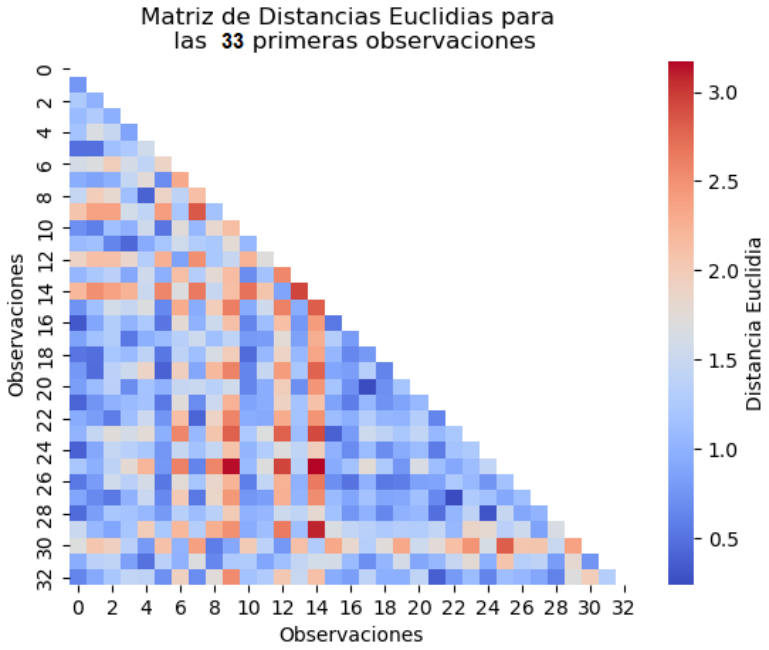

Figura 11: Matriz de distancias. Se ha eliminado el triángulo superior para eliminar contenido redundante.

Un análisis visual muestra algunas parejas de observaciones con distancias muy cercanas (5 y 0, 20 y 17, etc.) de las que se intuye que en un futuro podrán formar parte de un mismo clúster, mientras que otras se hallan notablemente alejadas (25 y 9, 29 y 14, etc.).

3.2. Determinación del Número de Clústeres - Análisis Jerárquico

Esta sección corresponde al apartado 2 de Clusterización de la hoja de ejercicios.

Una herramienta eficaz para identificar y visualizar patrones entre los diferentes elementos de un conjunto de datos es el dendrograma, una ordenación jerárquica de las observaciones tal que se agrupan en un mismo clúster aquellas observaciones con menor distancia entre ellas. La clusterización es jerárquica, esto es, las observaciones se ordenan sin tener un número de clústers prefijado. Si bien se pueden tener tantos clústers como uno quiera, es apropiado escoger aquel número que ofrezca una buena separación entre agrupaciones. Para ello, en la siguiente página se grafica el dendrograma que da cuenta de las relaciones entre las observaciones.

Se visualizan dos agrupaciones principales, una en verde y otra en naranja, que a su vez agrupan a muchas otras agrupaciones menores. Este fenómeno se basa en que las fusiones que componen la mayoría de los dos grupos ocurren a distancias relativamente cortas, indicando una mayor similitud dentro de cada grupo Ahora bien, se considera dicho número simplista e insuficiente para poder describir las agrupaciones reales de las que uno puede disponer, por lo que se escoge la formación de 3 grupos o clústers. Esta proposición se pone a prueba

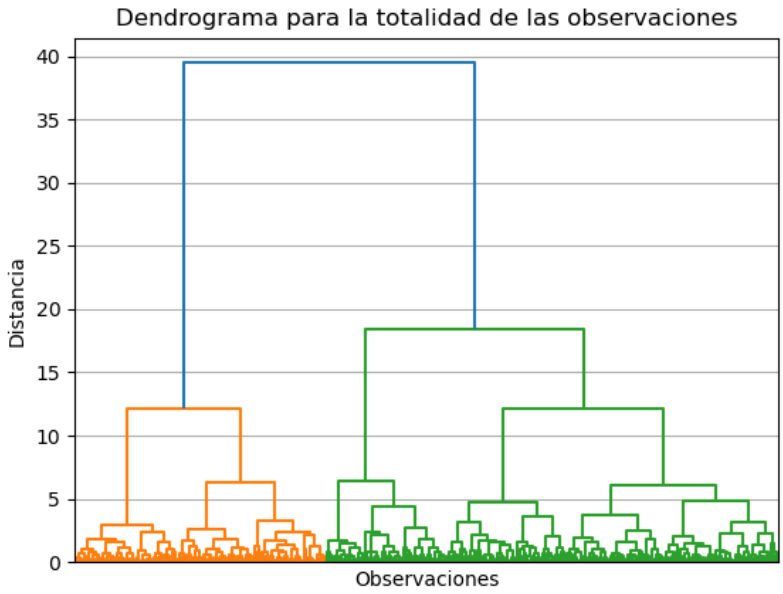

Figura 12: Dendrograma de la totalidad de las observaciones. En el eje horizontal se muestran las observaciones de donde se excluye el índice de las mismas para mejorar la legibilidad del gráfico. En el eje vertical se muestra la distancia a la que los clústers son fusionados durante el proceso de clusterización jerárquica.

con el *método del codo* para clusterizaciones. En este caso, se calcula el WCSS o Variabilidad dentro del conglomerado (suma de las distancias al cuadrado de cada observación respecto al centroide del grupo) dado una serie de números de clústers. A la par, se calcula el Puntaje de Silueta en función del número de clústers, una medida que indica para cada valor cuán similar es una observación a las otras en su mismo grupo en comparación con las observaciones de los clústers vecinos. El resultado es el siguiente:

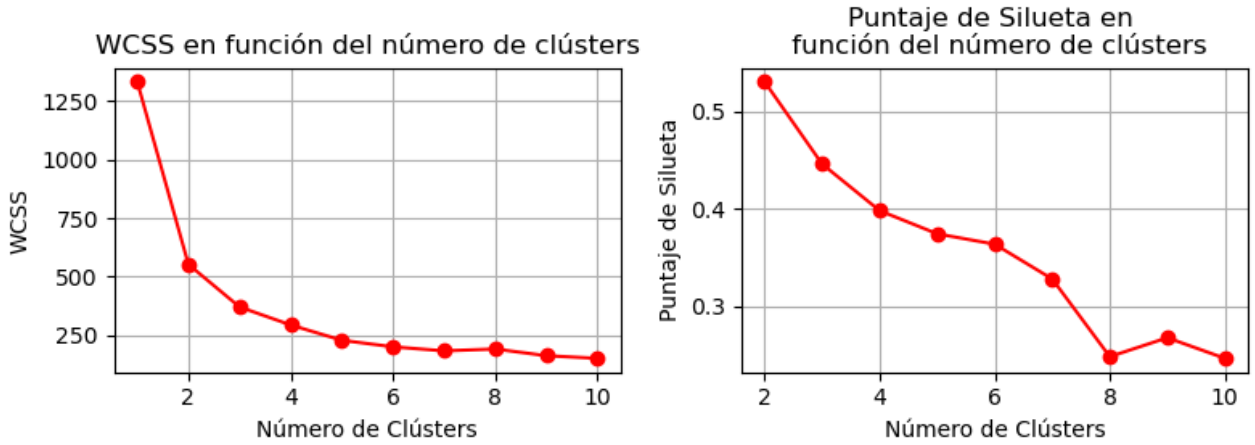

Figura 13: WCSS y Coeficiente de Silueta en función del número de clústers.

Se identifica un cambio de tendencia en el gráfico de WCSS a partir de los 2 clústers, por lo que se sugiere tomar 3 clústers. Este hecho se corrabora con el gráfico del Coeficiente de Silueta, que sugiere inicialmente tomar 2 clústers, pero considerando que esta elección podría ser demasiado simplista el siguiente valor más alto es el correspondiente a 3 clústers. Esto concuerda con el número de clústers que se ha considerado en el análisis jerárquico anterior. Alternativamente, se podrían utilizar 6 clústers dado el cambio de comportamiento del Coeficiente de Silueta, en contra del criterio de WCSS. Se desestima esta opción para no prolongar el presente documento excesivamente, si bien se comenta más abajo.

3.3. Análisis de Clúster No Jerárquico

Esta sección corresponde al apartado 3 de Clusterización de la hoja de ejercicios.

Una vez elegido el número de clústers óptimo mediante un análisis jerárquico, se muestra en función de las Componentes Principales escogidas las observaciones con su correspondiente etiqueta según a qué clúster pertenezcan, permitiendo solidificar la estructura de clústers identificada en la etapa jerárquica:

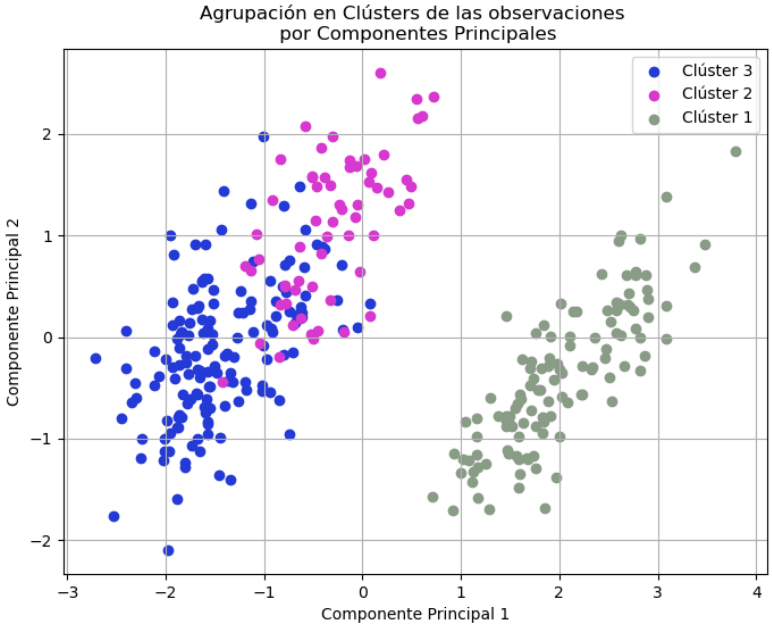

Figura 14: Especímenes en función del número de componentes principales con su respectiva etiqueta del clúster al que pertenecen.

Como uno puede observar, mediante *KMeans()* se identifican 3 agrupaciones, una aislada y dos solapadas que coinciden con las 3 diferentes especies de pingüinos. Esto refuerza la elección de haber tomado 3 clústers en vez de 2, ya que se habría tan solo podido identificar una única especie, y con dificultad las otras dos. Si se excluyesen los criteriors usados anteriormente de selección del número de clústers, uno puede indagar en posibles subestructuras del conjunto de datos, identificado aquellas agrupaciones de observaciones que se refieren a sexos opuestos o diferentes islas (de hecho, utilizando 6 clústers uno puede observar las agrupaciones que corresponden a las 3 especies y sus respectivos 2 sexos).

3.4. Evaluación de la Calidad de las Agrupaciones

Esta sección corresponde al apartado 4 de Clusterización de la hoja de ejercicios.

Se evalua la calidad de los clústers utilizado el Índice de Silueta, herramienta adelantada en la sección anterior. Permite mostrar la distribución de los Coeficientes de Silueta para cada clúster. Un coeficiente de silueta alto indica un mejor ajuste de la observación al clúster asociado, mientras que un valor bajo o negativo indica un mal ajuste. Se muestra a continuación el Índice de Silueta en función del clúster gráficamente:

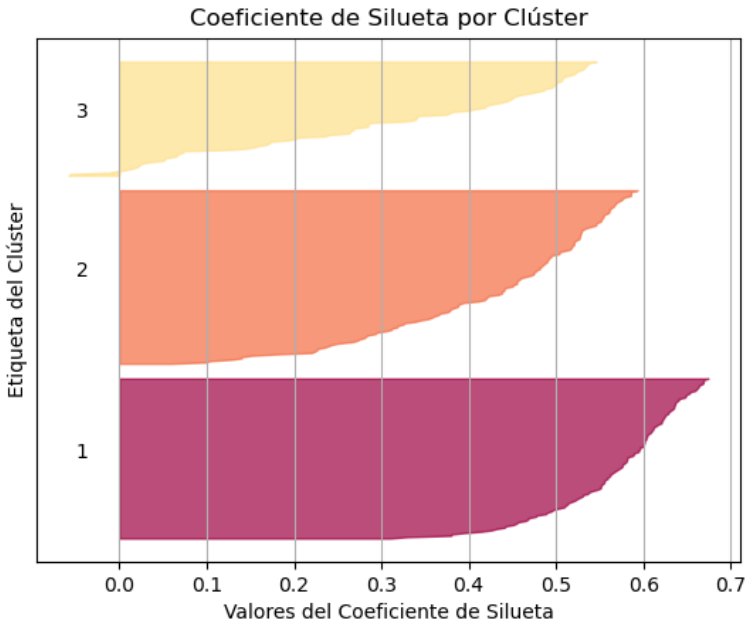

Figura 15: Coeficiente de Silueta para cada observación asociada a su respectivo clúster.

Se observa como el gráfico muestra altos valores del Coeficiente de Silueta para el clúster 1, aquel que se hallaba aislado de los 2 restantes. Por otra parte, el clúster 2 es el segundo grupo con mejores resultados, mientras que el 3 fracasa en algunas observaciones donde el Coeficiente de Silueta toma valores negativos. Este fenómeno es esperable ya que los clústers 2 y 3 se solapan, produciendo conflictos a la hora de escoger el clúster al que asociar la observación. Ahora bien, en general se han conseguido precisiones aceptables en la formación de los clústers.

3.5. Variables suplementarias

Esta sección corresponde al apartado 5 de Clusterización de la hoja de ejercicios.

El siguiente apartado pretende mostrar en un gráfico los centroides de cada especie y cada isla que tenga como ejes las dos primeras Componentes Principales. Esto permitirá discernir si existe alguna diferencia entre la fisionomía de los especímenes en función de la isla:

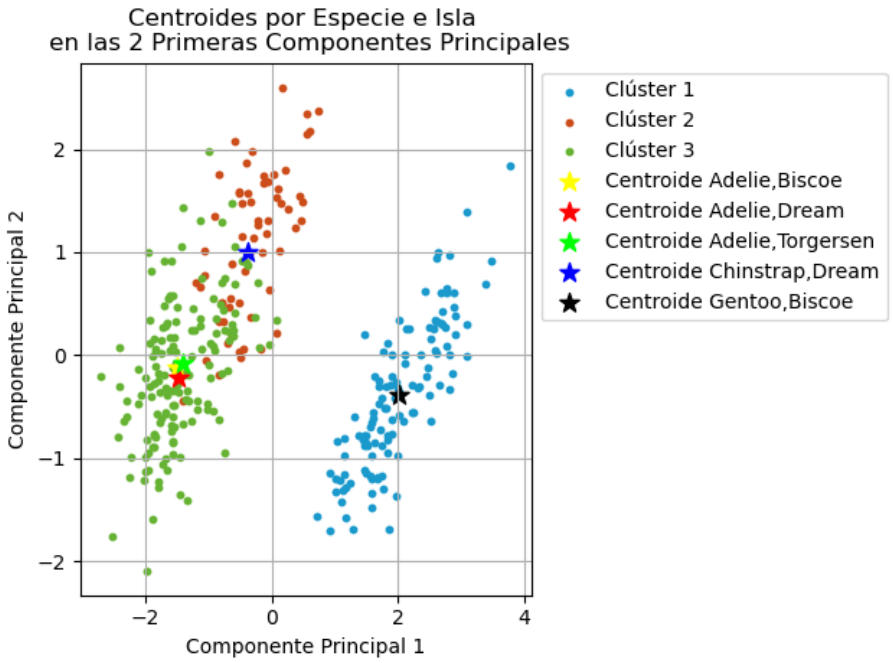

Figura 16: Observaciones en función de las 2 primeras Componentes Principales, etiquetadas según el clúster al que pertenecen. Se añaden a la vez los centroides de interés de algunas poblaciones particulares de especímenes.

En primer lugar, se observa como el centroide para la población de pingüinos Gentoo de la isla Biscoe cae en el Clúster 1, hecho que indica que dicho clúster corresponde a la población de dicha especie. En segundo lugar, el centroide de aquellos especímenes Chinstrap de la isla Dream se situa en el centro del segundo clúster, por lo que se reafirma que el segundo clúster corresponde a aquellos pingüinos de especie Chinstrap en la isla Dream. Por último, indiferentemente de la isla a la que se haga referencia, las tres poblaciones de pingüinos Adelie situan sus centroides en el mismo punto en el Clúster 3, por lo que se refuerzan una vez más las dos siguientes conclusiones: el Clúster 3 describe la población de pingüinos Adelie, y la isla donde se tomó la medida no es un factor influyente en las características fisionómicas de la especie Adelie.

3.6. Caracterización de los Clústeres

Esta sección corresponde al apartado 6 de Clusterización de la hoja de ejercicios.

Por último, se caracterizan los clústeres basándose en estadísticos descriptivos como la media y la desviación estándar. En la siguiente tabla se muestran las métricas descriptivas más relevantes:

Clúster	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
1	47.57±3.00	15.00±1.00	217.24±7.00	5092.44±500.00
2	49.75±2.50	18.60±1.10	197.11±7.00	3810.09±300.00
3	39.19±3.00	18.29±1.20	190.04±7.00	3680.10±500.00

El Clúster 1, de ahora en adelante, la población de pingüinos Gentoo, es la especie que presenta el mayor peso corporal (con la mayor dispersión de las 3 especies) y longitud de aletas (misma dispersión que para el resto de especies). En cambio, es la especie que presenta la menor profundidad de pico y la segunda mayor longitud de pico (ambas variables tienen dispersiones no muy alejadas de las correspondientes a las otras especies). Para las poblaciones de los Clústers 2 y 3 (especies Chinstrap y Adelie, respectivamente), se hallan valores medios del peso corporal, longitud de aleta y profundidad del pico muy similares, lo que genera conflictos a la hora de identificarlos correctamente. En efecto, en los gráficos anteriores se podía haber adelantado esta situación al tener que ambos clústers se solapan. En especial, la especie Adelie es la que posee un menor peso corporal (con la mayor desviación estándar relativa, donde esta representa cerca del 13 % de su media) y longitud de pico.

4. Conclusiones

Tras el análisis de los resultados obtenidos, se llega a las siguientes conclusiones:

1. El número adecuado de Componentes Principales es 2.
2. El número adecuado de clústers a realizar es 3.
3. La especie Gentoo posee una fisionomía que la diferencia claramente de las especies Chinstrap y Adelie.
4. Las especies Chinstrap y Adelie poseen características fisiológicas similares que dificultan la clusterización de los grupos.
5. Se identifican diferencias notables en la fisionomía del espécimen en función de su sexo.
6. Para la especie Adelie, no hay indicios suficientes para confirmar que la isla donde el espécimen habita influye en su fisionomía.

5. Bibliografía

Las fuentes consultadas a lo largo de la redacción del presente documento y los códigos en Python se adjuntan a continuación:

- Martín García, Daniel. (2024). Análisis Clustering, Universidad Complutense de Madrid.
- Martín García, Daniel. (2024). Análisis de Componentes Principales, Universidad Complutense de Madrid.