

Minería de datos y modelización predictiva

HUGO GÓMEZ SABUCEDO

hugogomezsabucedo@gmail.com

Máster Big Data, Data Science & Inteligencia Artificial

Curso 2024-2025

Universidad Complutense de Madrid

Índice

1. Introducción y análisis	3
1.1. Introducción	3
1.2. Correlaciones entre las variables	4
2. Análisis PCA	5
2.1. Análisis inicial	5
2.2. PCA sobre los datos estandarizados	6
3. Clustering	10
3.1. Agrupamiento jerárquico	10
3.2. Agrupamiento K-means	10
3.3. Validación del agrupamiento	10
3.4. Comparación entre agrupamiento jerárquico y K-means	10
4. Interpretación de los resultados	11

1. Introducción y análisis

1.1. Introducción

En este ejercicio analizaremos el conjunto de datos de penguins disponible en la librería seaborn de Python. Este conjunto de datos contiene observaciones de diferentes ejemplares de pingüinos, para cada uno de los cuales se recoge su especie ('Adelie', 'Chinstrap' o 'Gentoo') la isla en la que se encuentra ('Biscoe', 'Dream' y 'Torgensen'), la longitud de su pico (*bill_length_mm*, en milímetros), la profundidad de su pico (*bill_depth_mm*, también en milímetros), la longitud de su aleta (*flipper_length_mm*, en milímetros), su masa corporal (*body_mass_g*, en gramos), y su género (que puede ser 'Male' para macho, 'Female' para hembra, o 'NaN' si no está disponible). El objetivo de este ejercicio es aplicar las diferentes técnicas de reducción de dimensionalidad y de agrupamiento, para ver si se pueden definir grupos de pingüinos con características similares.

En primer lugar, observamos una pequeña muestra de los datos para comprobar su estructura. Haciendo un `.head(5)` vemos los 5 primeros registros, en donde se observan algunos detalles interesantes. Observamos que hay valores perdidos (NaN) no sólo en el género, sino que hay un pingüino que tiene todos los valores perdidos, para el que no tenemos ningún dato. Tendremos que analizar la mejor estrategia para tratar este caso.

species	island	bill_length_mm	...	flipper_length_mm	body_mass_g	sex
Adelie	Torgersen	39.1	...	181.0	3750.0	Male
Adelie	Torgersen	39.5	...	186.0	3800.0	Female
Adelie	Torgersen	40.3	...	195.0	3250.0	Female
Adelie	Torgersen	NaN	...	NaN	NaN	NaN
Adelie	Torgersen	36.7	...	193.0	3450.0	Female

Cuadro 1

En la figura 1 se puede observar un análisis descriptivo de las variables numéricas. Lo primero a destacar es que se observan 2 NaN en cada variable, lo cual nos podría indicar que tenemos, además del pingüino que vimos anteriormente, otro con todos los datos faltantes. Esto se refuerza viendo que el conteo de las variables es de 342, cuando en el conjunto de datos tenemos 344 pingüinos. Respecto a la distribución de los datos, vemos que las dos primeras variables son bastante simétricas, con coeficientes de asimetría próximos a 0. Para las otras dos variables, se observa una simetría positiva, un poco más pronunciada en el caso del peso, pero en general es una distribución también simétrica. Tampoco se observan valores atípicos en los datos.

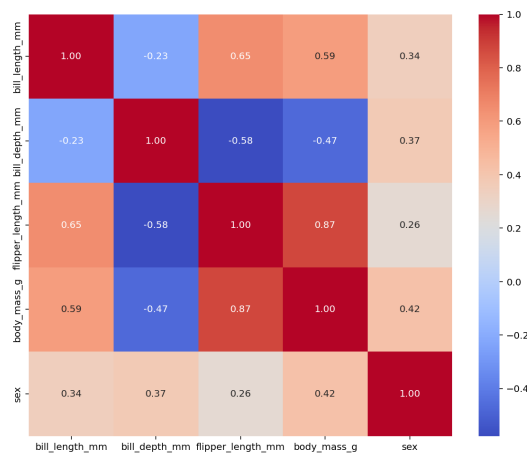
	count	mean	std	min	25%	50%	75%	max	Rango	Asimetría	Varianza	Coef. Variación	Missing
bill_length_mm	342.0	43.921930	5.459584	32.1	39.225	44.45	48.5	59.6	27.5	0.053118	29.807054	0.124302	2.0
bill_depth_mm	342.0	17.151170	1.974793	13.1	15.600	17.30	18.7	21.5	8.4	-0.143465	3.899808	0.115140	2.0
flipper_length_mm	342.0	200.915205	14.061714	172.0	190.000	197.00	213.0	231.0	59.0	0.345682	197.731792	0.069988	2.0
body_mass_g	342.0	4201.754386	801.954536	2700.0	3550.000	4050.00	4750.0	6300.0	3600.0	0.470329	643131.077327	0.190862	2.0

Figura 1: Análisis descriptivo de los datos

Respecto al tratamiento de los valores NaN, se han decidido eliminar por completo aquellos registros con algún valor missing, ya que la cantidad que se ha observado de los mismos es realmente pequeña. Como se comentó, se detectan dos ejemplares con todos los datos missing, además de nueve de ellos para los que falta el sexo. Es una proporción escasa de los datos, por lo que su eliminación no supondría una gran pérdida de información. Aún así, se recalca que, para el tratamiento de los valores missing, se podría aplicar algún algoritmo que los imputase manteniendo la proporción de género que hay originalmente. Se observa una distribución prácticamente simétrica, con 165 pingüinos hembra y 168 machos, por lo que, de los 9 pingüinos que tienen el sexo missing, podríamos establecer para cinco de ellos el sexo 'Female', y el sexo 'Male' para cuatro de ellos, manteniendo así el equilibrio.

Por último, convertiremos los valores de la variable sexo, ya que es dicotómica, mapeando el género macho como 1 y el género hembra como 0, de forma que podamos usar esta variable también en el análisis, proporcionándonos una nueva característica para los datos. De esta forma, sólo quedarían como variables categóricas la especie y la isla, las cuales se eliminarán también, ya que no tendría sentido mantenerlas. De todas formas, se guarda una copia del dataframe, ya que resultará útil de cara al análisis de los resultados ver en qué isla o de qué especie es cada pingüino y cuál predomina en cada grupo.

1.2. Correlaciones entre las variables



Si representamos la matriz de correlaciones, observamos una relación positiva fuerte entre el peso y la longitud de la aleta, así como entre la longitud del pico y el peso y la longitud de la aleta. Esto significa, por ejemplo, que los pingüinos que tienen mayor peso tienden a tener una aleta más larga. Sin embargo, entre el peso y la longitud de la aleta se observa una correlación negativa con la profundidad del pico, lo que nos indica que, cuanto más profundo es el pico, menos pesa y más corta tiene la aleta el pingüino. Cabe destacar también la relación negativa entre la longitud y la anchura del pico, lo cual indicaría que los pingüinos tienen picos anchos y cortos, o largos y estrechos.

Figura 2: Matriz de correlacion

2. Análisis PCA

2.1. Análisis inicial

Para comenzar con el análisis PCA o de componentes principales, lo primero que haremos será estandarizar los datos, ya que las escalas de las variables originales son distintas (unas unidades en mm, otras en gramos...), y uno de los requisitos para poder aplicar el PCA es que los datos estén en la misma escala o estandarizados. Como vimos en 1.1, los datos no presentan valores atípicos, y también se han limpiado de valores perdidos, por lo que los datos cumplen todos los requisitos para aplicar el PCA.

Creamos el PCA con cuatro componentes, el número máximo que se indica, y calculamos los autovalores, la varianza explicada y la varianza acumulada, recogiendo en la tabla 2.

	Autovalores	Varianza explicada	Varianza Acumulada
Componente 1	2.844160	0.567124	0.567124
Componente 2	1.411420	0.281436	0.848560
Componente 3	0.486707	0.097049	0.945609
Componente 4	0.172850	0.034466	0.980075

Cuadro 2: Tabla con autovalores

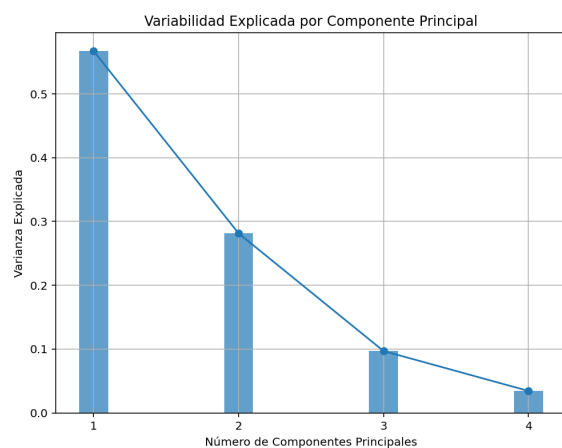


Figura 3: Variabilidad Explicada

Respecto a los autovalores, nos indican la cantidad de varianza explicada por cada uno de los componentes, de forma que cuanto mayor sea implica que esa componente retiene una gran cantidad de información de los datos originales. Estos valores son más elevados para los dos primeros componentes, que, además, si analizamos la varianza explicada (también en la figura 6) y la acumulada, vemos que entre los dos explican casi el 85 % de la variabilidad de las variables originales. Para elegir el número adecuado de componentes, tenemos en cuenta los siguientes criterios:

- Que los autovalores de las componentes sean mayores a 1. En este caso, sólo lo cumplen las componentes 1 y 2.
- Que los componentes expliquen al menos entre el 80 % y el 90 % de los datos originales (usando la varianza acumulada). Con las dos primeras componentes tenemos un 84.5 % de la varianza explicada.
- Observando el gráfico del codo (figura 6).

Por lo tanto, aplicando todos estos criterios, y viendo que en la tercera componente del gráfico del codo ya no se aprecia una ganancia significativa de información (contribuye apenas en un 10 %), se determina que el **número adecuado de componentes es 2**.

2.2. PCA sobre los datos estandarizados

Una vez que hemos determinado el número de componentes adecuados como 2, repetimos el PCA con este número de componentes principales. Obtenemos los autovectores para cada componente, así como las correlaciones entre cada componente y las variables originales (para ver que variables tienen más peso en cada componente), y el cuadrado de esta correlación (para ver qué proporción de la varianza de esa variable es explicada por ese componente). Los resultados obtenidos se recogen en la tabla 3. Sus representaciones gráficas se encuentran englobadas en la figura 8.

Variable	Autovector 1	Autovector 2	Corr. C1	Corr. C2	\cos^2 C1	\cos^2 C2
bill_length_mm_z	0.462289	0.168869	0.778463	0.200320	0.606004	0.040128
bill_depth_mm_z	-0.331708	0.652723	-0.558574	0.774291	0.312005	0.599527
flipper_length_mm_z	0.563897	-0.094552	0.949563	-0.112162	0.901669	0.012580
body_mass_g_z	0.554251	0.047941	0.933320	0.056869	0.871085	0.003234
sex_z	0.226018	0.730888	0.380598	0.867013	0.144855	0.751712

Cuadro 3: Métricas del PCA

En lo que respecta a la interpretación de los resultados, por un lado tenemos los autovectores, que indican la contribución de cada variable original a una componente en concreto. Esto lo vemos también en la figura 4, donde vemos que las variables que más contribuyen en el autovector 1 son 'flipper_length_mm_z', 'body_mass_g_z' y 'bill_length_mm_z', mientras que 'sex_z' y 'bill_depth_mm_z' tienen una menor contribución. Lo contrario ocurre en el caso del autovector 2. Apoyándonos también en la figura 5, esto nos sugiere que la primera componente está relacionada con el tamaño en general del pingüino (longitud de la aleta y masa, así como longitud de pico), mientras que la segunda componente captura más información sobre las diferencias entre los pingüinos machos y hembra. De esta forma, las especies de pingüinos más grandes tomarán también valores más grandes en la componente 1, mientras que en la componente 2 tomarán valores más altos aquellas especies con picos más profundos o que tengan diferencias más marcadas entre los machos y las hembras. También observamos la varianza explicada de cada variable (figura 6) y el gráfico de dispersión de las observaciones en base a cada variable (figura 7).

Además, en la figura 9, podemos ver el mismo gráfico de la figura 7, pero con cada observación coloreada en función de la especie a la que pertenece, y con una forma distinta en

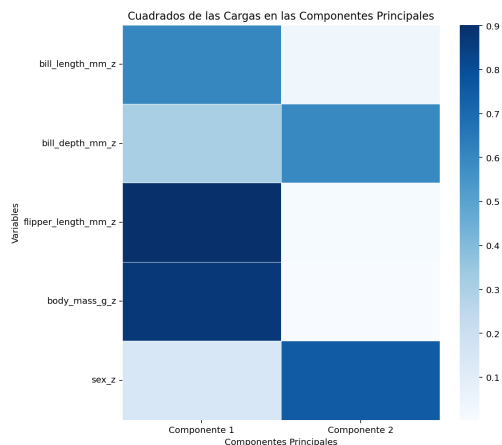


Figura 4: Cargas en las componentes principales

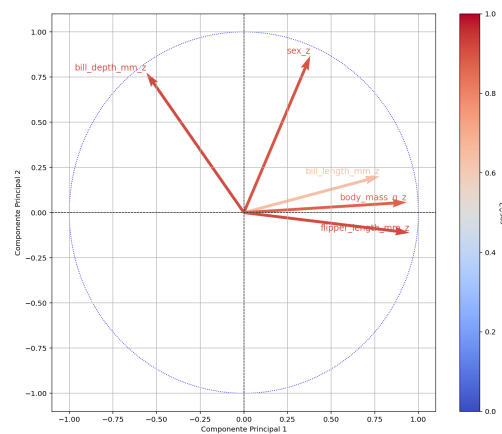


Figura 5: Autovectores para componentes principales

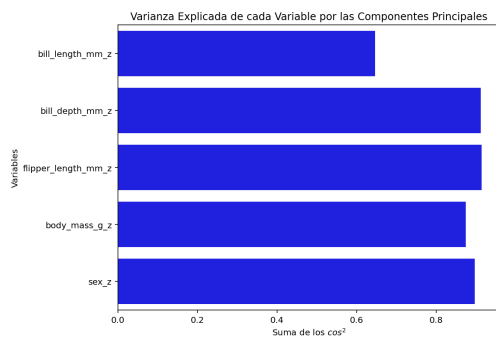


Figura 6: Varianza explicada de cada variable

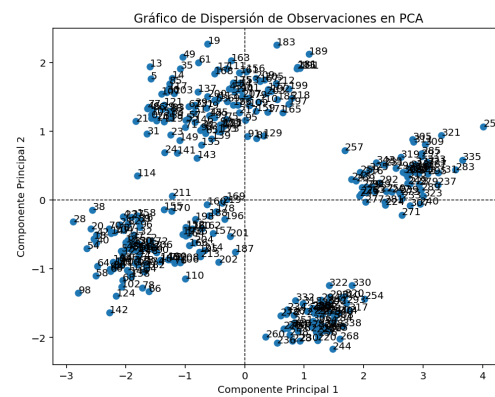


Figura 7: Gráfico de dispersión de las observaciones

Figura 8: Conjunto de gráficas

función del género del pingüino (y eliminando también el índice de cada pingüino, para hacer el gráfico más legible y que se aprecien más las diferencias entre especies). El código de este gráfico, que parte de la función `plot_pca_scatter`, puede observarse en 1.

Inicialmente, observamos cuatro grupos diferenciados, dos de los cuales se corresponden exclusivamente con la especie *Gentoo*. Para esta especie, vemos que todos sus individuos presentan un comportamiento similar para la componente 1 (que, como dijimos, está relacionada con el tamaño de los pingüinos). Además, curiosamente, los dos subgrupos que se observan se encuentran justo en lados opuestos de la componente 2, que era la que definimos como el marcador de género, y podemos observar que todos los *Gentoo* hembra toman valores de la componente 2 positivos, mientras que los *Gentoo* macho toman valores negativos. Esto nos sugiere nuestras suposiciones iniciales sobre lo que representaba cada componente.

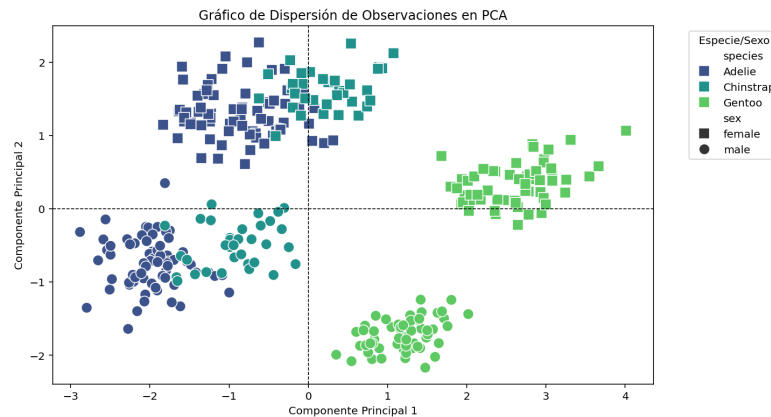


Figura 9: Gráfico de dispersión por especie

Además, para las especies *Adelie* y *Chinstrap* se observa una situación parecida: ambas toman valores similares para la componente 1, lo que nos sugiere que son individuos con características físicas y de peso similares; y están polarizadas, aunque con menos diferencia que los *Gentoo*, en función del género del pingüino. Estas especies son las que destacan menos en la primera componente, o las que destacan más "negativamente" (indicando, quizás, individuos de menor tamaño que los *Gentoo*).

```

1  penguins_std['species'] = penguins_original['species']
2  penguins_std['sex'] = penguins_original['sex'].map({0: 'male', 1: 'female'})
3
4  for i in range(n_components):
5      for j in range(i + 1, n_components): # Evitar pares duplicados
6          # Crear el gráfico de dispersión
7          plt.figure(figsize=(10, 6)) # Ajusta el tamaño de la figura si es necesario
8
9          # Colorear los puntos según la especie y cambiar la forma según el sexo
10         sns.scatterplot(
11             x=componentes_principales[:, i],
12             y=componentes_principales[:, j],
13             hue=penguins_std['species'], # Colorear por especie
14             style=penguins_std['sex'], # Cambiar forma según el sexo
15             palette='viridis', # Paleta de colores para las especies
16             s=100, # Tamaño de los puntos
17             markers={'male': 'o', 'female': 's'} # Marcadores para machos y hembras
18         )
19
20         # Dibujar líneas discontinuas que representen los ejes
21         plt.axhline(0, color='black', linestyle='--', linewidth=0.8)
22         plt.axvline(0, color='black', linestyle='--', linewidth=0.8)

```



```

23
24     # Etiquetar los ejes
25     plt.xlabel(f'Componente Principal {i + 1}')
26     plt.ylabel(f'Componente Principal {j + 1}')
27
28     # Establecer el título del gráfico
29     plt.title(f'Gráfico de Dispersión de Observaciones en PCA')
30
31     # Mostrar la leyenda
32     plt.legend(title='Especie/Sexo', bbox_to_anchor=(1.05, 1),
33               loc='upper left')
34
35     plt.show()

```

Listing 1: Código para la obtención del gráfico de dispersión

Para valorar conjuntamente las características físicas de los pingüinos, podemos construir un índice que considere ambas componentes. Una primera aproximación podría ser construirlo en base a una combinación ponderada de las dos componentes principales, basándose en el \cos^2 que tiene cada uno de las componentes (que vimos en 2.1), de forma que tendríamos algo como $ndice = w1 * CP1 + w2 * CP2$ (donde $w1$ y $w2$ son la suma del \cos^2 para $CP1$ y $CP2$).

Para calcular el índice para, por ejemplo, Adelie, podemos emplear el siguiente código, donde se calcula en primer lugar los valores de las componentes principales para esta especie, para a continuación aplicar el índice que hemos calculado. De esta forma, obtenemos un valor de -3.653316 para el índice, lo que nos indica que los pingüinos de esta especie tienden a tener valores bajos en ambas componentes, lo que sugiere que son más pequeños en tamaño, con picos menos profundos, y menos diferencia entre géneros.

Si repetimos el proceso para Chinstrap, obtenemos un índice de -0.161348, y para Gentoo de 4.574419. Si observamos la gráfica 9, podemos corroborar esto, ya que vemos que los pingüinos más pequeños son claramente los Adelie, con un tamaño similar a los Chinstrap, pero con estos últimos presentando un mayor marcaje de género; mientras que los Gentoo son tanto los más grandes como los que más marcaje de género presentan.

```

1     w1 = cos2['Componente 1'].sum() # Peso para CP1
2     w2 = cos2['Componente 2'].sum() # Peso para CP2
3     resultados_pca = pd.DataFrame(componentes_principales, columns=['
4     Componente 1', 'Componente 2'])
5     resultados_pca['Índice'] = (w1 * resultados_pca['Componente 1']) + (
6     w2 * resultados_pca['Componente 2'])
7
8     adelie_data = penguins_std[penguins_std['species'] == 'Adelie']
9     adelie_cp = pca.transform(adelie_data.drop(columns=['species', 'sex'
10    ]))
11     adelie_cp_df = pd.DataFrame(adelie_cp, columns=['Componente 1', '
12     Componente 2'])
13     adelie_cp_df['Índice'] = (w1 * adelie_cp_df['Componente 1']) + (w2 *
14     adelie_cp_df['Componente 2'])
15     print(f"Índice Adelie: {adelie_cp_df['Índice'].mean()}")

```

3. Clustering

3.1. Agrupamiento jerárquico

3.2. Agrupamiento K-means

3.3. Validación del agrupamiento

3.4. Comparación entre agrupamiento jerárquico y K-means

4. Interpretación de los resultados