PCA en Blue Jays

Waldir Toscano, Mausel Perez, Jorge Acosta

22/03/2025

Resumen

Se realizó un análisis de componentes principales (PCA) en un dataset de medidas morfológicas de aves *Blue Jays* para reducir la dimensionalidad. Los resultados indican que 4 componentes principales explican el 93.32 % de la varianza total, superando el umbral del 90 %. Las variables más influyentes fueron *body_ mass_ g*, *skull_ size_ mm*, y dimensiones del pico (*bill_ depth_ mm*, *bill_ width_ mm*). Este análisis permite simplificar futuros modelos sin perder información crítica.

1. Introducción

El PCA es una técnica estadística utilizada para reducir la dimensionalidad de conjuntos de datos complejos, identificando patrones subyacentes. En biología, es especialmente útil para analizar medidas morfológicas en especies animales. Este estudio aplica PCA a un dataset de 123 muestras de *Blue Jays* con 7 variables numéricas, con el objetivo de determinar cuántos componentes son necesarios para explicar el 90 % de la varianza e interpretar su significado biológico.

2. Objetivos

Los objetivos de esta actividad son:

- Utilizar PCA para disminuir las dimensiones y realizar un análisis sobre el dataset Blue Jays.
- Realizar analisis bivariado todos contra todos.
- Interpretar resultados del análisis bivariado.
- Calcular la matriz de covarianza y sus valores y vectores propios.
- Determinar cuantos y cuales componentes son necesarios para describir el 90 % de la varianza de los datos.

3. Descripción de la Actividad

Se reutiliza el código entregado por la institución, con el fin de realizar un análisis al dataset Blue Jays, esto utilizando la clase(PCA) que se encontraba implementada en dicho código. Se realiza un código extra para previamente ajustar los datos y realizar los objetivos de la actividad.

3.1. Codificación

Listing 1: Se cargan los datos se limpian se modifica la columna sex para reemplazar M y F por valores numericos y posteriormente se genera un pairplot con seaborn para realizar una analisis

```
# Cargar datos
df = pd.read_csv('/content/blue_jays.csv')
df_numerico = df.drop(['bird_id', 'sex'], axis=1)
target = df['sex'].replace({'M': 0, 'F': 1})  # Codificar sexo para colorear


# Analisis bivariado
sns.pairplot(df, hue='sex', vars=df_numerico.columns)
plt.show()
```

Listing 2: Se utiliza la clase PCA entregada por la institucion para aplicarle PCA a los datos cargados y modificados anteriormente por ultimo se obtienen el valor propio y se calculan la varianza explicada y acumulada

```
# Aplicar PCA
data = df_numerico.to_numpy()
pca = PCA(n_componentes=4)
datos_pca = pca.run(data)

# Varianza explicada
valores_propios = pca.valores_propios
varianza_explicada = valores_propios / np.sum(valores_propios)
varianza_acumulada = np.cumsum(varianza_explicada)
```

Listing 3: Por ultimo se imprimen los resultados anteriores y se dibuja la grafica

```

#Se imprimen los resultados en la consola
print("Varianza explicada por componente (%):")
print(np.round(varianza_explicada * 100, 2))
print("Varianza acumulada (%):")
print(np.round(varianza_acumulada * 100, 2))

# Grafico de componentes
pca.dibujar("Componentes Principales (Blue Jays)", df_numerico
    .columns, target, datos_pca)
```

4. Gráficas

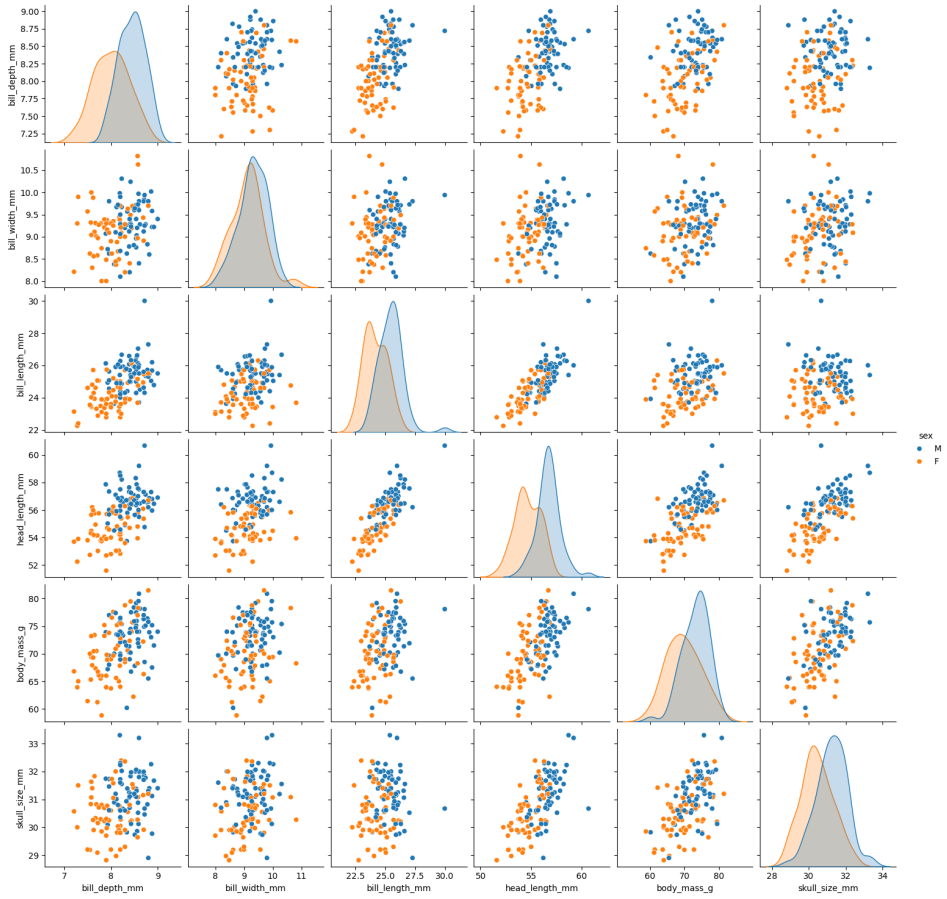

Figura 1: Pairplot Blue Jays

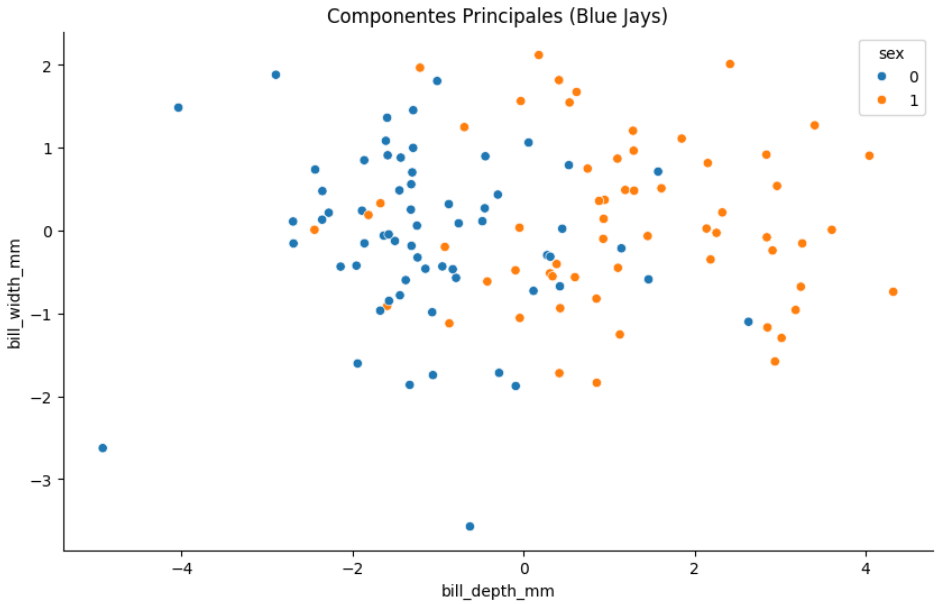

Figura 2: PCA Blue Jays

5. Análisis de gráficas

Figura 1. En esta gráfica se observa que el tamaño general de las aves (pico, cabeza, cuerpo, cráneo) está interrelacionado. Aves con picos más grandes tienden a tener cabezas y cuerpos más grandes, y viceversa, por lo que esto nos puede indicar adaptaciones evolutivas para alimentarse de los recursos que tengan disponible en su zona. También en algunos casos, se observa una clara separación entre machos y hembras, lo que nos puede indicar que dependiendo el sexo pueden tener un ligero cambio o diferencia en sus cuerpos.

Figura 2. En esta gráfica se puede confirmar lo mencionado sobre el sexo de las aves en la gráfica 1, ya que se observa que los machos suelen ser mas grandes que las hembras. Ademas las diferencias en la forma del pico podrían estar vinculadas a la especialización en el consumo de recursos. Por ejemplo, picos anchos y profundos son útiles para romper semillas, mientras que picos delgados son eficaces para capturar insectos. Esto podría dar a entender la adaptación de estas aves a su zona, dependiendo de los recursos de que dispongan en esta misma.

6. Interpretación

6.1. Varianza Explicada

Componente	Varianza Explicada (%)	Varianza Acumulada (%)
PC1	54.10	54.10
PC2	16.72	70.83
PC3	13.45	84.28
PC4	9.04	**93.32**

6.2. Relación con Variables Originales

- **PC1 (54.1 %)**: Asociado a *body_ mass_ g* y *skull_ size_ mm* (tamaño corporal). Los machos (M) dominan este componente.
- **PC2 (16.72 %)**: Vinculado a *bill_ depth_ mm* y *bill_ width_ mm* (forma del pico).
- **PC3 (13.45 %)**: Relacionado con *head_ length_ m* (tamaño craneal).
- **PC4 (9.04 %)**: Captura variabilidad residual de características menores.

6.3. Conclusiones

En base a los datos anteriores se puede determinar lo siguiente:

- Se requieren **4 componentes principales** para describir el 93.32 % de la varianza, cumpliendo el objetivo del 90 %.
- Las variables más críticas son las asociadas al tamaño corporal (*body_ mass_ g*) y craneal (*skull_ size_ mm*), seguidas de medidas del pico.
- La reducción a 4 componentes simplifica el dataset, facilitando análisis posteriores (clustering, clasificación por sexo) sin perder información relevante.
- Los machos (M) presentan valores significativamente mayores en PC1, reflejando dimorfismo sexual en tamaño corporal.

7. Conclusión general

Los análisis realizados revelan dos patrones clave en la morfología de las Blue Jays. Primero, existe una fuerte correlación entre el tamaño del pico, la cabeza, el cuerpo y el cráneo, sugiriendo que estas aves han desarrollado proporciones corporales integradas como adaptación a sus recursos alimentarios. Por ejemplo, individuos con picos más grandes y robustos, asociados a cuerpos más voluminosos, podrían estar especializados en consumir alimentos duros como semillas,

mientras que aquellos con estructuras más delgadas podrían priorizar insectos u otros recursos.

Segundo, la separación observada entre machos y hembras en ciertas variables, como la masa corporal y el tamaño craneal, apunta a un dimorfismo sexual moderado. Esto podría reflejar presiones evolutivas diferenciadas: los machos, al ser ligeramente más grandes, podrían tener ventajas en competencias por territorio o parejas, mientras que las hembras podrían optimizar su morfología para roles como la incubación o la búsqueda eficiente de alimento.

Estos hallazgos subrayan cómo la forma y el tamaño corporal en las Blue Jays no son aleatorios, sino el resultado de adaptaciones a su entorno ecológico y a dinámicas sociales intrínsecas a su especie. El uso combinado de gráficos bivariados y PCA permite no solo visualizar estas tendencias, sino también cuantificar su impacto en la variabilidad total de la población, ofreciendo una ventana a los mecanismos evolutivos que moldean su diversidad morfológica.