Taller 1 de PCA

Sebastián Cotes, Yefferson González, Lorann Peñuela

IES, INFOTEP Instituto Nacional de Formación Técnica Profesional "HVG"

Humberto Velazquez Garcia

Abstract

En esta actividad, se nos pide realizar un código que tome los datos de un archivo CSV llamado Blue Jays, el cual trata de un estudio realizado a las aves de esa especie. Con los datos ya listo realizamos una codificación de un PCA que analizara los datos y nos proveerá de información relevante del estudio.

1 Introducción

PCA a tomado gran importancia a lo largo de los años, este es un método estadístico que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales. El objetivo principal de PCA es reducir la dimensionalidad de los datos mientras se conserva la mayor cantidad posible de variabilidad presente en los datos originales. En este caso los datos sera proporcionados por un archivo facilitada por nuestro profesor el cual se titula "Blue Jays", con este archivo se implementará PCA y se realizara un análisis a sus resultados.

2 Objetivos

Los objetivos principales de esta actividad son:

- Implementar PCA para el análisis del archivo Blue Jays.
- Calcular la matriz de covarianza para entender cómo las variables se relacionan entre sí.
- Calcular los valores propios y vectores propios.
- Seleccionar los componentes principales que explican la mayor parte de la variabilidad en los datos en un 90 porciento.

3 Descripción de la actividad

Se nos pide realizar un código que tome los datos de un archivo CSV llamado Blue Jays, el cual trata de un estudio realizado a las aves de esa especie. Con los datos ya listo realizamos una codificación de un PCA, este es un método

estadístico que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales y poder que analizar los datos y que nos proveerá de información relevante del estudio.

1. Utilizar dataset blue jays.
2. Realizar análisis bivariado todos contra todos.
3. Interpretar resultados del análisis bivariado.
4. Calcule la matriz de covarianza y sus valores y vectores propios.
5. Determine cuantos y cuales componentes son necesarios para describir el 90 porciento de la varianza de los datos.

3.1 Codificación

Se utilizaron varias librerías para realizar esta actividad. Solo se proporcionará la codificación realizada por los participantes, mas no el proporcionado por la institución.

Listing 1: Llamamos el documento csv blue jays y lo asignamos a una varible

```
blue_jays = pd.read_csv("/content/blue_jays.csv")
data = blue_jays
target = blue_jays[data.columns[1]]
```

Listing 2: Creamos las graficas correspondientes al DF de todos contra todos y analizamos resultados

```
sns.pairplot(blue_jays)
plt.show()
```

Listing 3: Limpiamos los datos y creamos una categorizacion para la columna "sex" con valores de 1 para masculino y 0 para femenino. Empezamos la creacion matriz de covarianza y sus vectores.

```
data = data.drop(columns=['bird_id'])
data['sex'] = data['sex'].map({'M': 1, 'F': 0})
pca = PCA()
datos_pca = pca.run(data)

print("% de varianza descrita por cada caracteristica o valor
    propio")
print((pca.valores_propios/np.sum(pca.valores_propios))*100)

#Para describir el 90% de la varianza se necesitan 5
    componentes
np.cumsum((pca.valores_propios/np.sum(pca.valores_propios))
    *100)

# Imprimir la matriz de covarianza
pca.calc_matriz_covarianza()

# Dibuja las graficas segun los datos proporsionados.
pca.dibujar("Blue Jays",[data.columns[0], data.columns[1]],
    target, datos_pca)
```

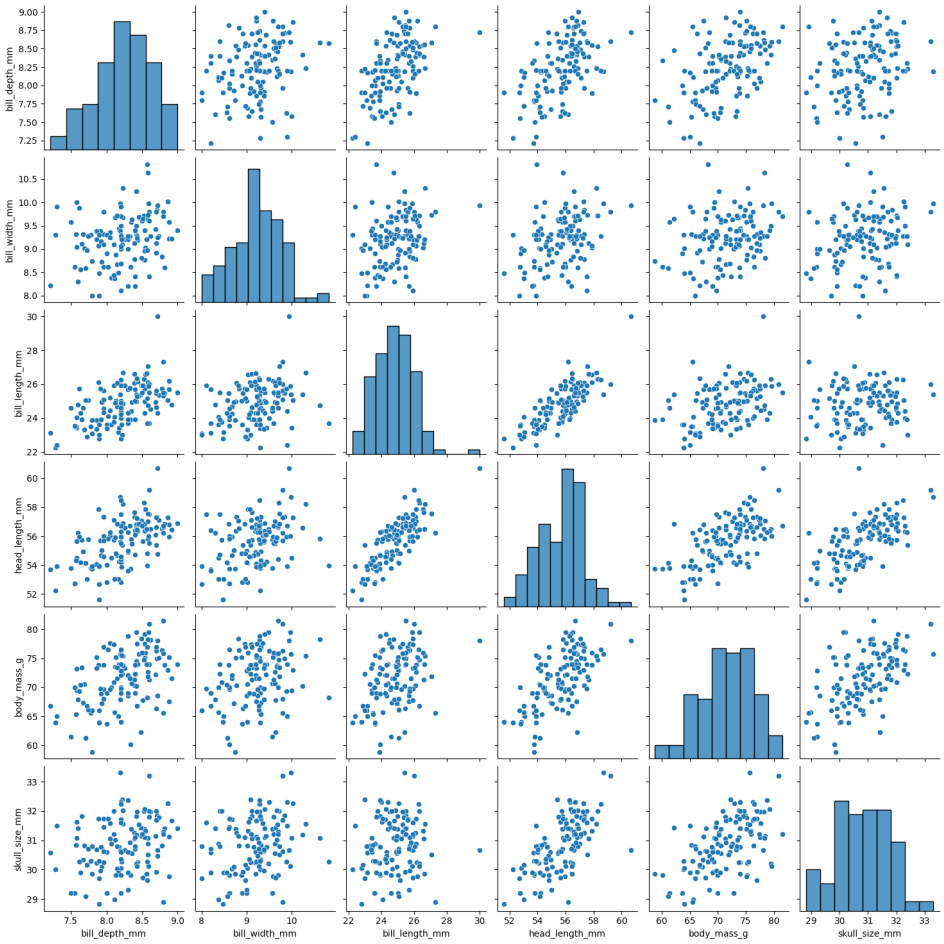

Figure 1: **Gráfica de todos contra todos**

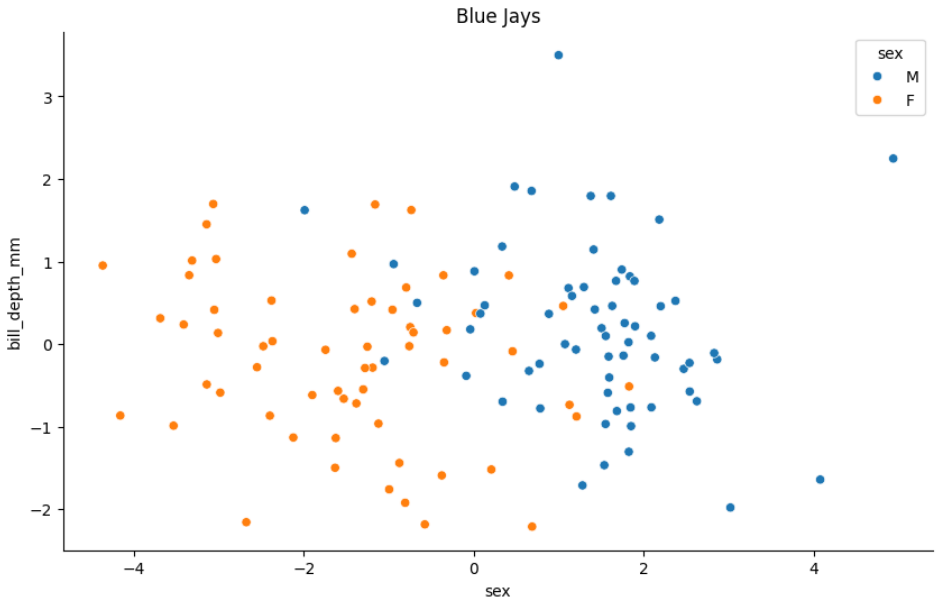

Figure 2: **Comparación ya con el PCA aplicado**

4 Análisis de las gráficas

Figura 1. En esta figura se realiza la comparacion de todods contra todos, solo se tomaran algunas de las graficas y se realizara la inter pretacion.

Petal Length vs Petal Width: Este gráfico muestra cómo el largo se relaciona con el ancho. Una correlación positiva indicaría que a medida que el largo de las alas aumenta, el ancho también tiende a aumentar.

Sepal Length vs Petal Length: Este gráfico muestra la relación entre el Sepal Length y el Petal Length. Con una correlación positiva indica que a medida que el Sepal Length aumenta, el Petal Length también tiende a aumentar.

Además de las relaciones específicas mencionadas, pueden observarse correlaciones generales entre todas las variables. Esto ayuda a identificar patrones consistentes y hacer predicciones basadas en las relaciones entre las variables.

5 Conclusión

El análisis de componentes principales (PCA) nos ayudó a reducir la dimensionalidad del conjunto de datos, manteniendo la mayor parte de la variabilidad presente en los datos originales. Las primeras componentes principales capturan la mayor parte de la variabilidad, permitiendo una visualización más sencilla y un análisis más eficiente. La visualización de las dos primeras componentes principales muestra cómo se distribuyen las muestras en función de las componentes principales y la variable objetivo. La categorización de la columna de sexo y la inclusión de la variable objetivo en el análisis PCA nos permitió entender mejor las relaciones entre las variables y cómo se agrupan las muestras.