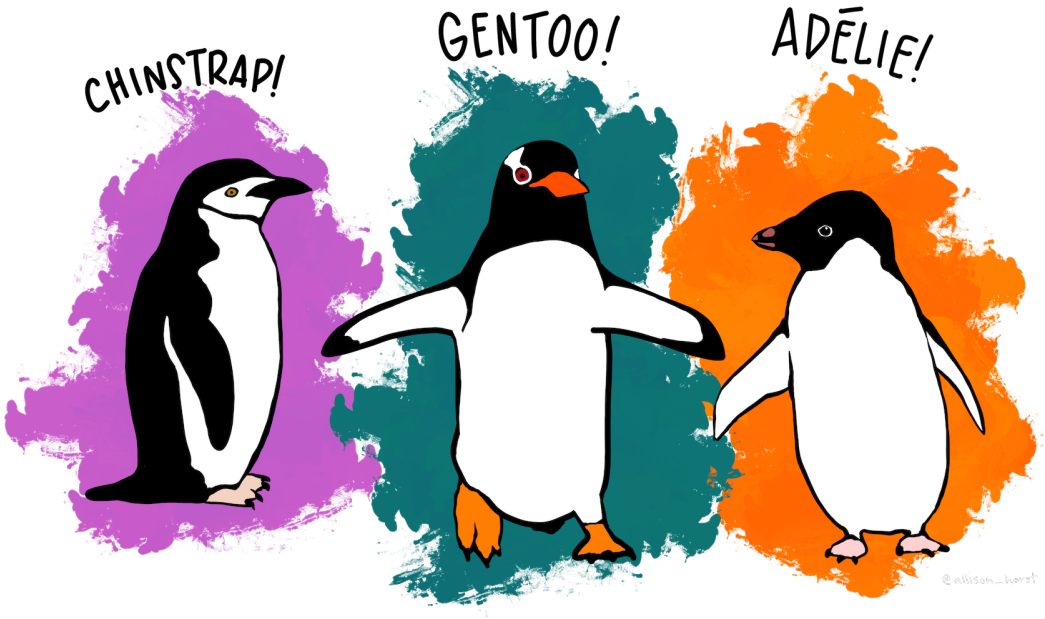
Entrega 3

María Inés Fariello

2022-06-28



Palmerpenguins

En este trabajo vamos a usar los datos de los pingüinos de la estación de Palmer recolectados por Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Las figuras fueron creadas por @allison_horst, éstas y los datos están disponibles en su página de github .

Para anlizar este juego de datos comenzaremos descargando los datos y elminando los que tengan datos faltantes en alguna columna que refiera a las medidas tomadas de los pingüinos. Estos datos ya limpios están disponibles en la página del curso.

```
library(palmerpenguins)
library(ggplot2)
library(tidyverse)
library(GGally)
library(dplyr)
library(tree)
```

```
data(penguins)
penguinsLimpio <- penguins[which(is.na(penguins$bill_length_mm)==FALSE),]
```

Para facilitarles el trabajo, veamos las variables que contiene este juego de datos y cómo se relacionan entre ellas.

```
penguinsLimpio %>%
    dplyr::select(species, body_mass_g, ends_with("_mm")) %>%
    GGally::ggpairs(aes(color = species)) +
    scale_colour_manual(values = c("darkorange","purple","cyan4")) +
    scale_fill_manual(values = c("darkorange","purple","cyan4"))
```

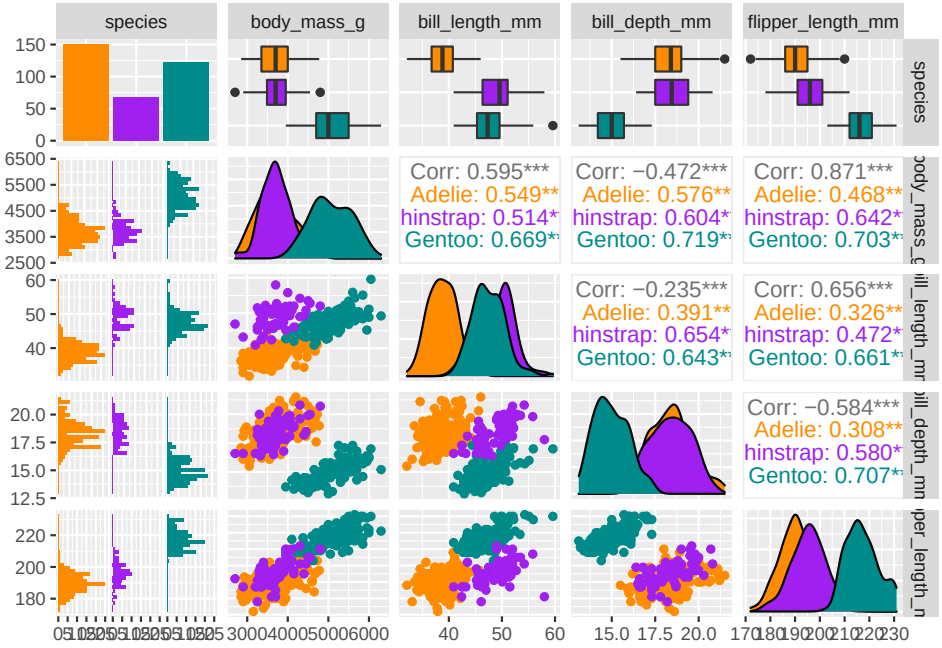
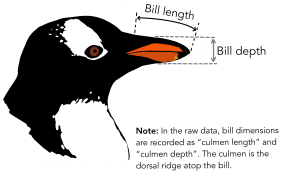


Figure 1: Relaciones entre las variables medidas de los Pingüinos de la estación Palmer

De esta figura, deducimos que probablemente las dimensiones del pico sean importantes para clasificar a estas especies. En la figura siguiente, vemos a que se refiere cada una de las medidas del pico.



Manos a la obra: Clasificación en las especies de los pingüinos

Clasificación de Gentoo vs el resto utilizando solamente una variable.

1) Crear una variable que valga 1 si el pingüino es Gentoo, 0 sino y separar o crear una variable que permita dividir los datos en un conjunto de entrenamiento y otro de test a partir de los índices proporcionados en la sección Datos del curso.

2) A partir de la figura 1 discutir si la variable `bill_depth_mm` puede ser una variable adecuada para separar Gentoo del resto. ¿Hay alguna otra variable que le parezca adecuada para esta tarea?

Regresión logística

3) Utilizando una regresión logística, predecir si el pingüino es Gentoo o no, basándose solamente en `bill_depth_mm` (usar como límite de probabilidad 0.5).

4) a) Calcular el porcentaje de error total del modelo en la muestra train. Identificar los individuos en los que se equivoca el modelo e investigar si tienen alguna característica particular, que explique de alguna manera por qué se está equivocando el modelo al predecir.
   b) Repetir la parte 4a) con el modelo ajustado para la muestra de entrenamiento, para la muestra test.

Análisis discriminante en una dimensión

5) Suponiendo que ambos grupos (Gentoo y no Gentoo) tienen la misma varianza, encontrar el límte entre las dos distribuciones a partir de un análisis discriminante utilizando la variable `bill_depth_mm`. Este límite debe calcularse "a mano" a partir de los parámetros necesarios, que se calcularán a partir de los datos.
6) Calcular los errores de clasificación en la muestra de entrenamiento y de test y comparar con los resultados obtenidos en la sección anterior (partes 4a y 4b).

Clasificación en las tres clases a partir de dos variables.

En esta sección utilizaremos las variables `bill_length_mm` y `flipper_length_mm` para clasificar las tres especies de pinguinos usando Análisis discriminante lineal y Árboles de decisión.

7) Graficar `bill_length_mm` en función de `flipper_length_mm` utilizando diferentes colores para cada especie.

8) Deducir las ecuaciones de las rectas que separan a las tres especies usando las ecuaciones del discriminante lineal.Graficar cómo queda dividido el espacio, a partir de las rectas calculadas (puede realizarse a mano).

Para ello supondremos que los tres grupos tienen la misma varianza. A continuación se presenta el cálculo de la matriz de varianza covarianza Σ y el resultado obtenido. (No es necesario que ustedes realicen el cálculo ni escriban el código, aunque lo pueden hacer si es que lo precieren.)

```
set.seed(12234)
indicesTrain<- sort(sample(dim(penguinsLimpio)[1],274, replace=FALSE))
penguinsTrain<-penguinsLimpio[indicesTrain,]

penguinsTrain <- penguinsTrain %>%
  dplyr::group_by(species) %>%
  dplyr::mutate(bill_length_c=scale(bill_length_mm,center=TRUE, scale=FALSE), flipper_length_c=scale(fli

(sigma<-cov(penguinsTrain[,c("flipper_length_c","bill_length_c")]))
##                  flipper_length_c bill_length_c
## flipper_length_c        42.651738      8.547301
## bill_length_c            8.547301      8.521055
```

9) Ajustar un árbol de clasificación para predecir la especie a partir de las variables `flipper_length_c`,`bill_length_c`. Dibujar en la gráfica las fronteras determinadas por los árboles (esta gráfica también puede realizarse a mano).

10) Calcular los errores de clasificación en las muestras de entrenamiento y test para ambos modelos y discutir cuál de los modelos le parece más adecuado para la tarea de clasificación de ambas especies.

Clasificación en las tres clases a partir de las variables y sexo.

11) Ajustar un árbol de clasificación a partir de todas las variables de medidas (pico, ala y peso) y el sexo. Calcular los errores de predicción en el conjunto de entrenamiento y de test. Comparar con los modelos obtenidos en la sección anterior.