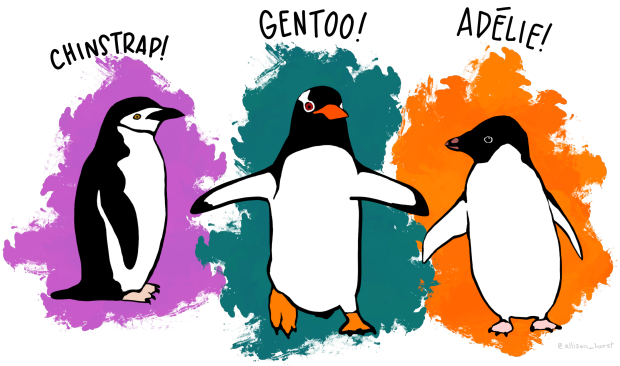
TP2 - Reconocimiento de Patrones

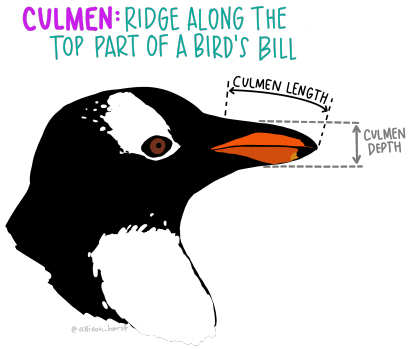
Elio Campitelli

1 Datos

En este TP voy a estar usando la base de datos `penguins` del paquete de R **palmerpenguins** (KB, TD, and WR 2014). Los datos recolectados por la Dra. Kristen Gorman en la Estación Palmer, consisten en mediciones de la longitud del culmen[1], alto del culmen y la masa corporal de 342 pingüinos de las especies *Pygoscelis adeliae* (Pingüino de Adelaida), *Pygoscelis papua* (Pingüino Juanito), y *Pygoscelis antarcticus* (Pingüino barbijo).



[1] El culmen es la parte superior del pico de las aves.



En la Tabla 1 se muestran las primeras 3 mediciones para cada especie.

Table 1: Primeras 3 entradas de cada especie en los datos utilizados

Especie	Longitud de culmen [mm]	Alto de culmen [mm]	Longitud de la aleta [mm]	Masa corporal [g]
adelaida	39.1	18.7	181	3750
adelaida	39.5	17.4	186	3800
adelaida	40.3	18.0	195	3250
juanito	46.1	13.2	211	4500
juanito	50.0	16.3	230	5700
juanito	48.7	14.1	210	4450
barbijo	46.5	17.9	192	3500
barbijo	50.0	19.5	196	3900
barbijo	51.3	19.2	193	3650

La variable categórica a predecir va a ser la especie, y las posibles variables predictoras son las dimensiones del culmen, la longitud de

la aleta y la masa corporal. Es decir, en principio es un espacio de dimensión 4. Como la idea es trabajar en $\mathbb{R}^2$, conviene explorar qué grado de separación permite cada combinación de dos variables. Esto se muestra en la Figura 1 donde se grafican scatterplots para todas las combinaciones de dos variables con la especie representada con color. Mirando las densidades de probabilidad (gráficos en la diagonal) se puede ver que la longitud del culmen separa bastante bien entre pingüino de adelaida y el resto mientras que las otras variables separan bien al pingüino juanito. Por lo tanto, las combinaciones que incluyen la longitud del culmen (gráficos en la primera columna) separan bastante bien entre las tres especies, mientras que el resto de las combinaciones tienen algún grado de mezcla entre pingüino de adelaida y pingüino juanito.

Figure 1: Scatteplot de todas las combinaciones de variables posibles en $\mathbb{R}^2$. En la diagonal, estimaciones de densidad de cada variable separadas por especies.

Para hacer las cosas más interesantes, voy a seleccionar el espacio formdado por el largo de la aleta y la masa. En este espacio $\mathbb{R}^2$, los piingüinos juanitos se separan muy bien de los otros, pero los adelaida y los barbijo están mezclados y son imposibles de separar.

La Figura 2 muestra distintas particiones posibles del espacio a partir de estimar las densidades de probabilidad conjunta de cada clase y asignando la clase con mayor densidad de probabilidad en

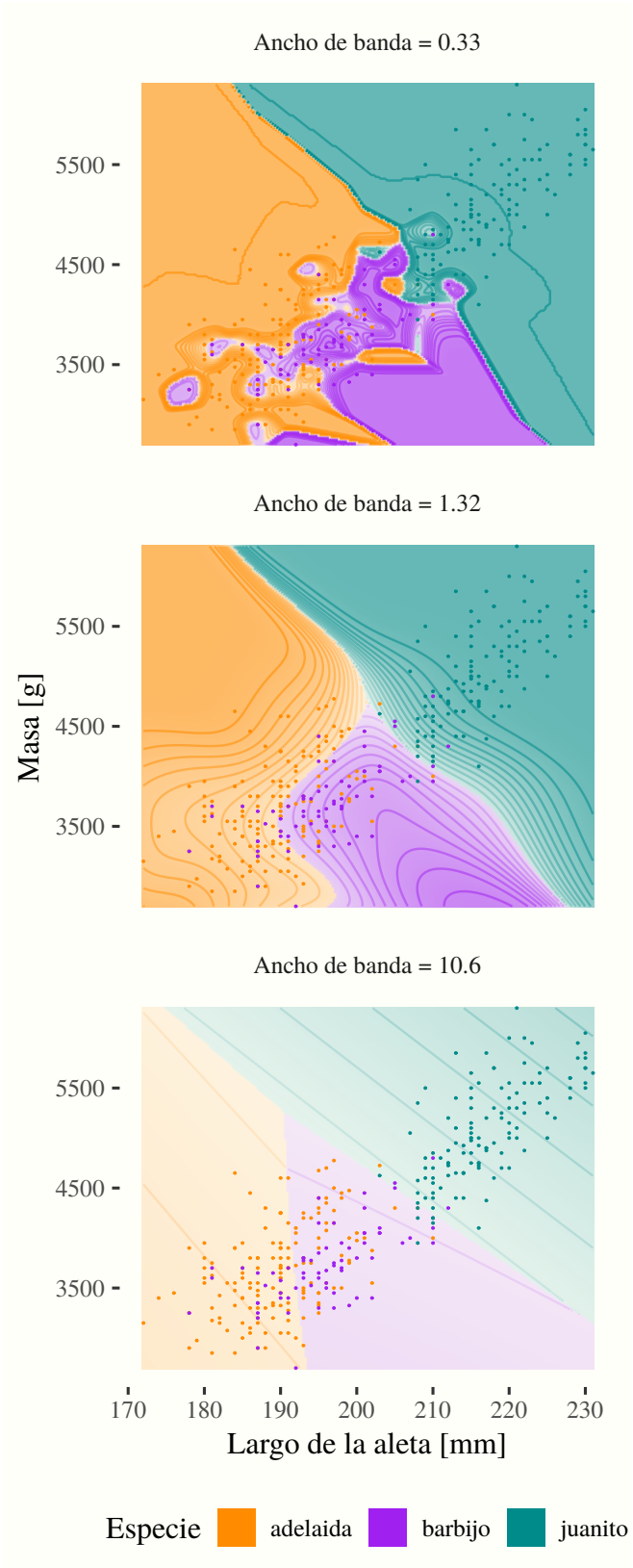


Figure 2: Separación del espacio a partir de estimar la densidad con un kernel gausiano de distintos anchos de banda.

cada punto[2]. Cuando el ancho de banda es muy pequeño, se observa que el clasificación sufre de overfitting. Las regiones clasifican perfectamente los datos observados, pero los límites de decisión se curva para rodear puntos aislados. Para un ancho de banda muy grande, los límites de decisión tidneden a rectas y se ve que las estimaciones de densidad son círculos concéntricos que ignoran la covarianza entre los datos.

[2] Dada la diferencia de escalas de x e y, en principio el valor del ancho de banda no podría ser el mismo para ambas dimensiones. Para armonizarlas, se hicieron los cálculos en base a las variables estandarizadas

2 *Clasificador cuadrático*

La función `clasificador_cuadratico` genera un modelo lineal de clasificación que es esencialmente un modelo lineal multivariado donde las K variables dependientes representan a las K categorías usando one-hot encoding. Es decir, el modelo tiene M predictores y K predicciones, una para cada clase. La clasificación se hace asignando la clase que tiene el valor máximo. Como medida de la confianza del resultado, se toma la razón entre ese valor máximo y la suma de todos los valores asignados a cada clase)

La partición del espacio $\mathbb{R}^2$ elegido usando el resultado del clasificador cuadrático se muestra en la Figura 3. Dado que el clasificador es lineal, las divisiones entre categorías son rectas que se interceptan en un punto central.

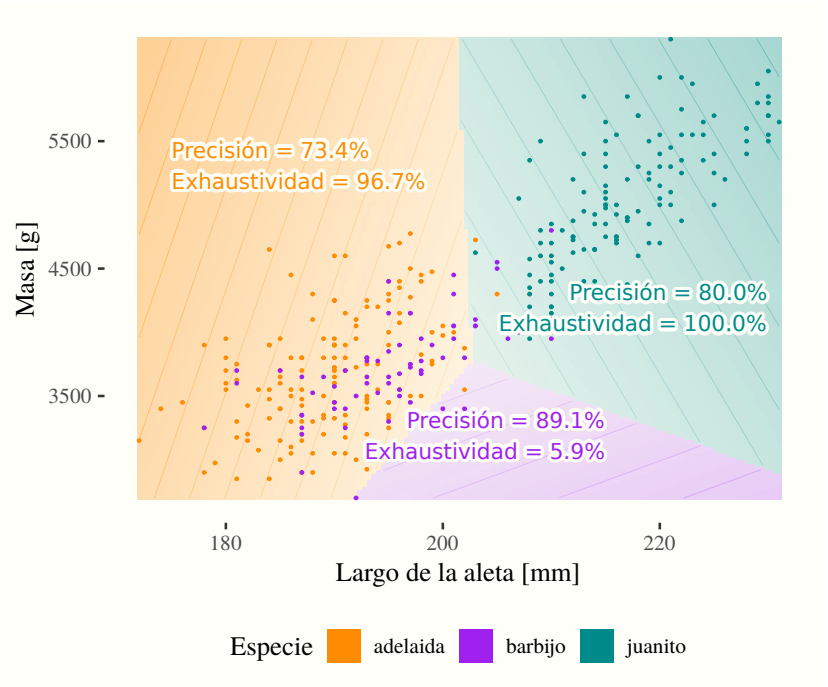


Figure 3: Clasificación en base a clasificador cuadrático lineal. En contornos negros, el nivel de confianza del modelo. "Precisión" se define como la proporción de observaciones clasificadas como una determinada especie que fueron clasificadas correctamente, "Exhaustividad" se define como la proporción de observaciones de cada especie correctamente clasificadas.

Se muestran dos medidas de la clasificación para cada especie. "Exhaustividad" es la proporción de observaciones que son clasificadas como una especie de forma correcta. "Precisión" es la proporción de observaciones de una determinada especie que son clasificadas correctamente. Es decir, la exhaustividad del 100% para los pingüinos juanitos implica que la probabilidad de que un pingüino juanito sea correctamente clasificado es del 100%. Sin embargo, la precisión del 80% implica que si el modelo clasifica un pingüino como juanito, hay un 80% de probabilidad de que haya sido clasificado correctamente. Se puede ver que si bien el clasificador lineal cuadrático identifica sin problemas a los pingüinos juanito, la separación entre adelaida y barbijo no es para nada buena.

2.1 Logística

La función `logistica` realiza el ajuste logístico mediante un método iterativo. Soporta clasificación de múltiples clases usando el algoritmo de uno-vs-todos. Es decir, para K clases genera K modelos que dan la probabilidad de que una observación determinada pertenezca a la clase k-ésima o a cualquiera de las otras. La clasificación luego se hace tomando la clase que tiene la mayor probabilidad. Al igual que con el clasificador cuadrático, como medida de confianza del modelo se calcula la razón entre la probabilidad asignada a la clase ganadora y la suma de todas las probabilidades asignadas.

La Figura 4 muestra el resultado de la clasificación logística lineal. Al igual que con la Figura 3, los límites de decisión son rectas dado qu eel modelo es linear en este espacio $\mathbb{R}^2$. La partición no es muy distinta de la partición usando el clasificador cuadrático. La Figura 5, en cambio, muestra el resultado del modelo logístico pero aplicado al espacio de dimensión 4 (x, y, x^2, y^2, xy). Al tener términos no lineales, los límites de decisión ahora pueden ser curvos[3]

[3] Son curvos en el espacio $\mathbb{R}^2$ mostrado, en el espacio de dimensión 4, siguen siendo rectas.

3 Expectation Maximisation

Expectation maximisation (EM) es una técnica no supervisada, es decir, que no tiene en cuenta las clases observadads.

La Figura 6 muestra los resultados del algoritmo de EM con los datos de pingüinos con 2 y 3 clases respectivamente. Los puntos son las medias de las distribuciones normales y las elipses marcan la el cuartil del 95%. Escala de colores de las clases de EM y las 3 especies no son las mismmas ya que no hay concordancia entre ambas necesariamente. El espacio se particiona entre las clases a partir de la densidad de probabilidad normal de cada clase encontrada por EM y tomando la clase que maximiza la misma.

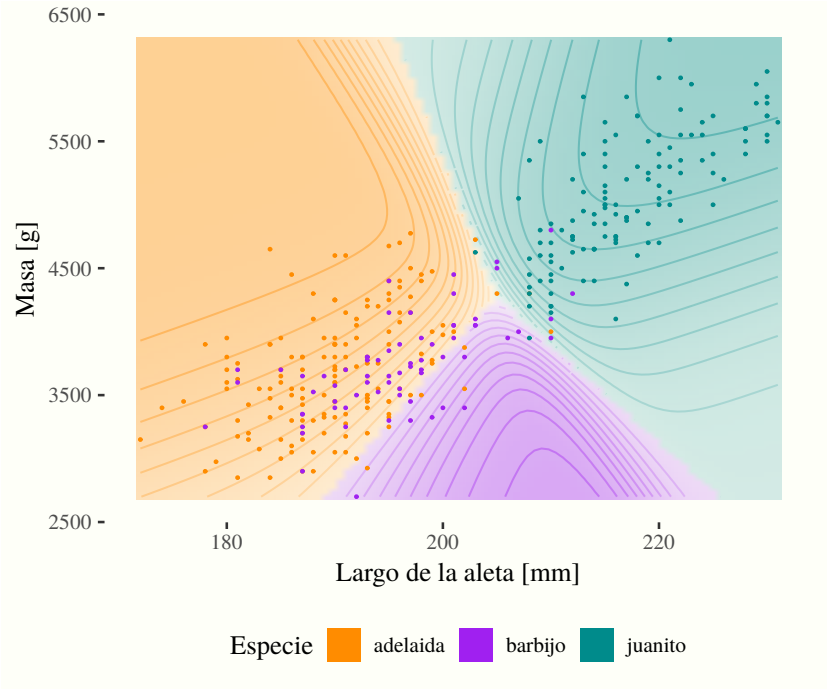


Figure 4: Convergencia del modelo logístico lineal. En contornos negros, el nivel de confianza del modelo.

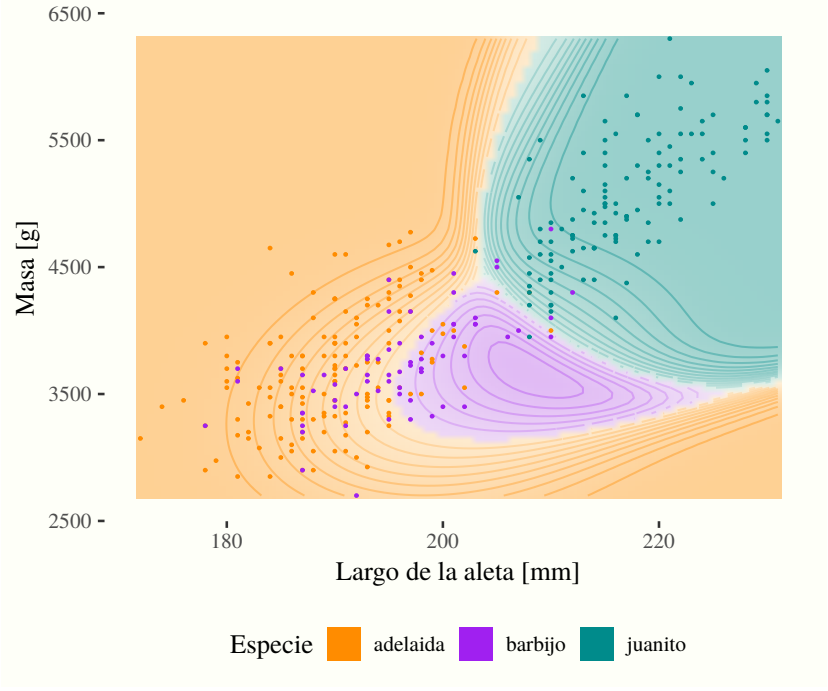


Figure 5: Igual que la Figura pero para un ajuste elíptico.

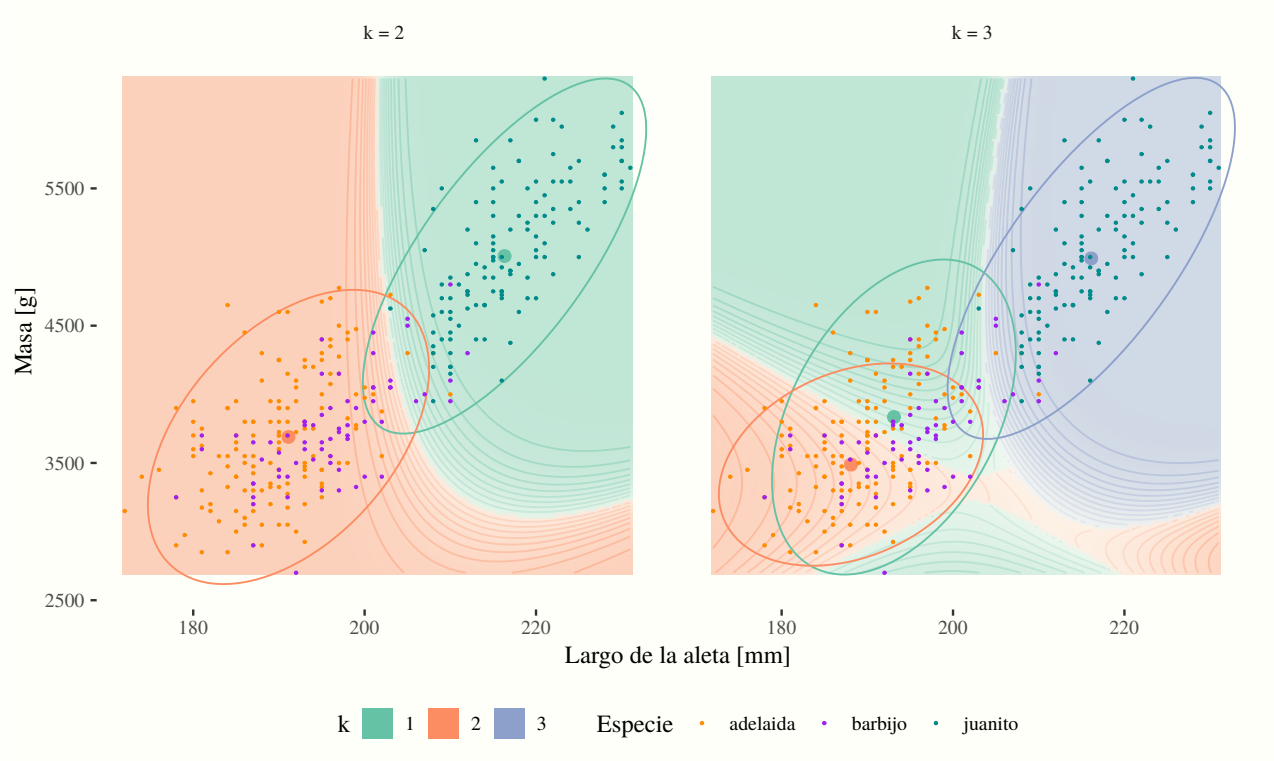


Figure 6: Medias y elipses normales representando el cuantil del 95% para los clusters identificados por Expectation Maximisation con k = 2 (izquierda) y k = 3 (derecha).

La parciticón con 2 clases separa correctmaente los pingüinos juanito de las otras dos especies. La partición con 3 clases, en cambio, identifica el *cluster* de los pingüinos juanito, pero no hay separación entre los otras dos especies.

Referencias

KB, Gorman, Williams TD, and Fraser WR. 2014. "Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis)." *PLoS ONE* 9(3) (e90081): –13. `https://doi.org/10.1371/journal.pone.0090081`.