Trabajo Práctico N°2 - Modelos de Clasificación

Aprendizaje Estadístico - FIUBA

1er Cuatrimestre - 2020

José F. González - 100063 - jfgonzalez@fi.uba.ar

Atento a: Ing. Jemina García Revisión 1.8

1. Introducción

Nos interesa construir un modelo de clasificación para predecir si un espécimen de abulone[1] es adulto o infante dados su `longitud`, `peso.total` y `anillos`. Para entrenar el modelo de clasificación disponemos de 4177 observaciones de abulones ya clasificados con sus distintas medidas, de las cuales definimos el subconjunto data.tr con el 80 % (3341 obsevaciones) para entrenar los distintos modelos y el subconjunto data.te con el restante 20 % para evaluarlos y compararlos. **Como criterio general buscamos minimizar la tasa de error en los conjuntos de entrenamiento, maximizando la medida *accuracy***. Siendo que hay suficientes datos disponibles para entrenar y evaluar no se usaran métodos de remuestreo.

2. Modelos de Regresión Logística

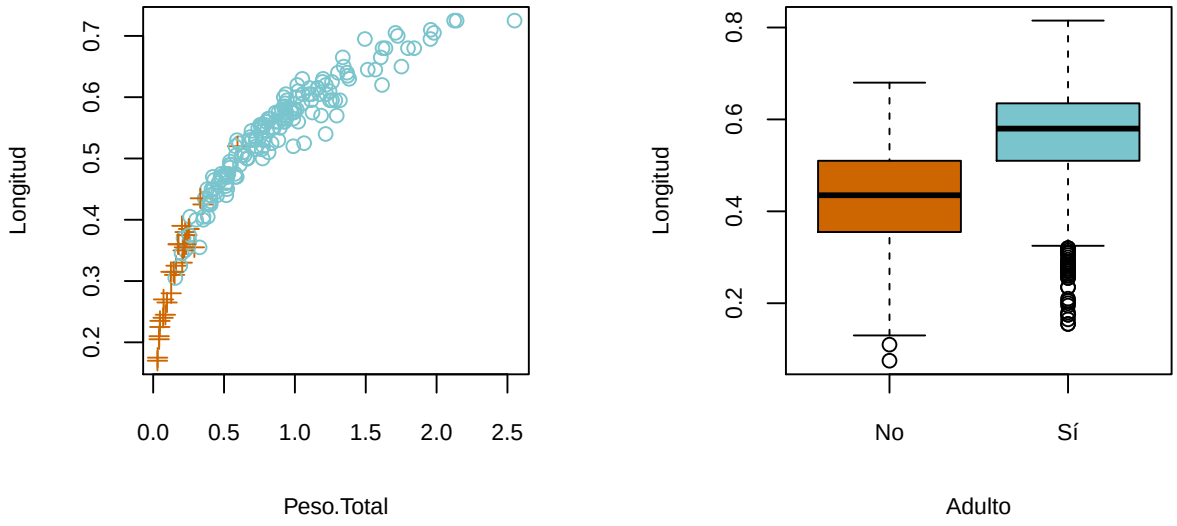
2.1. Primer Modelo Logístico - Clasificación por longitud

Comenzamos proponiendo distintos modelos logísticos y comparandolos entre ellos. Estos modelos serán distintas formas de estimar la probabilidad de que la variable `adulto` tome uno entre dos valores `Sí` o `No`.

En la Figura 1 se graficaron la `longitud` y `peso.total` para el subconjunto `data.tr`. En el panel izquierdo de la Figura 1 se muestran los primeros doscientos casos de especímenes adultos en celeste y los infantes en naranja . Parece que los casos infantes tienden a tener menor longitud que los adultos. En el panel derecho de la Figura 1 se muestra el diagrama de cajas de la distribución de la `longitud` partida en la variable binaria adulto.

El primer modelo logístico que analizamos es el más simple, queremos estimar la probabilidad `adulto` dado `longitud`. Para ello utilizamos la función logística de la Ecuación 1 donde los parámetros β_0 y β_1 se estiman por máxima verosimilitud y se muestran en el Cuadro 1.

Figura 1: **Izquierda:** La longitud y peso total de doscientos especímenes. Los especímenes adultos se muestran en celeste, los infantes en naranja. **Derecha:** Boxplot de la longitud como función de la varaible binaria `adulto` que toma valores `Sí` o `No`.



[1] Haliotis - Wikipedia

$$P(\texttt{adulto} = \texttt{Sí} | \texttt{longitud}) = \frac{e^{\beta_0 + \beta_1 \texttt{longitud}}}{1 + e^{\beta_0 + \beta_1 \texttt{longitud}}} \tag{1}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.35	0.23	-22.99	0.00
longitud	12.25	0.46	26.36	0.00

Cuadro 1: Coeficientes de regresión logística para la probabilidad de un espécimen sea adulto dada su longitud estimados con los datos de data.tr

El Cuadro 1 muestra las estimaciones de coeficientes para el primer modelo. Vemos que $\hat{\beta}_1 = 12{,}25$ indicando que un incremento en `longitud` de 0,1 aumenta un 1,2 el $log(odds)$ equivalente a un aumento de 0,016 en la probabilidad de ser `adulto`. Los p-valores del Cuadro XXX están asociados al test con hipótesis nula $H_0 : \beta_1 = 0$ y la rechazan a un nivel de significación $\alpha << 1$, es decir, hay suficiente evidencia de una asociación entre `longitud` y la probabilidad de `adulto`. Utilizando estas estimaciones el modelo nos dice que, por ejemplo, la probabilidad estimada de que un espécimen sea adulto cuando su longitud 0,4 será

$$\hat{p} = \frac{1}{1 + e^{5{,}35 - 12{,}25 \times 0{,}4}} = 0{,}39 \tag{2}$$

Antes de comenzar a evaluar el modelo con los datos de prueba se debe elegir un umbral de decisión. Utilizando la curva ROC de la Figura 2 se usa el umbral de probabilidad $p = 0{,}74$ para separar entre adulto e infante. Ahora, utilizando el conjunto `data.te` evaluamos al modelo sobre datos distintos a los de entrenamiento. Estimando las probabilidades y utilizando el corte $p = 0{,}74$ se construye la matriz de confusión del Cuadro **??**. En ella se ve que se predicen correctamente 417 adultos y 232 infantes, una tasa global de 78 %. Mientras que se clasifican incorrectamente 116 adultos como infantes y 71 infantes como adultos, dando una tasa de error de 22 %. Resumimos el desempeño de este clasificador sobre las métricas *accuracy*, *precision* y *recall* que se muestran en el Cuadro

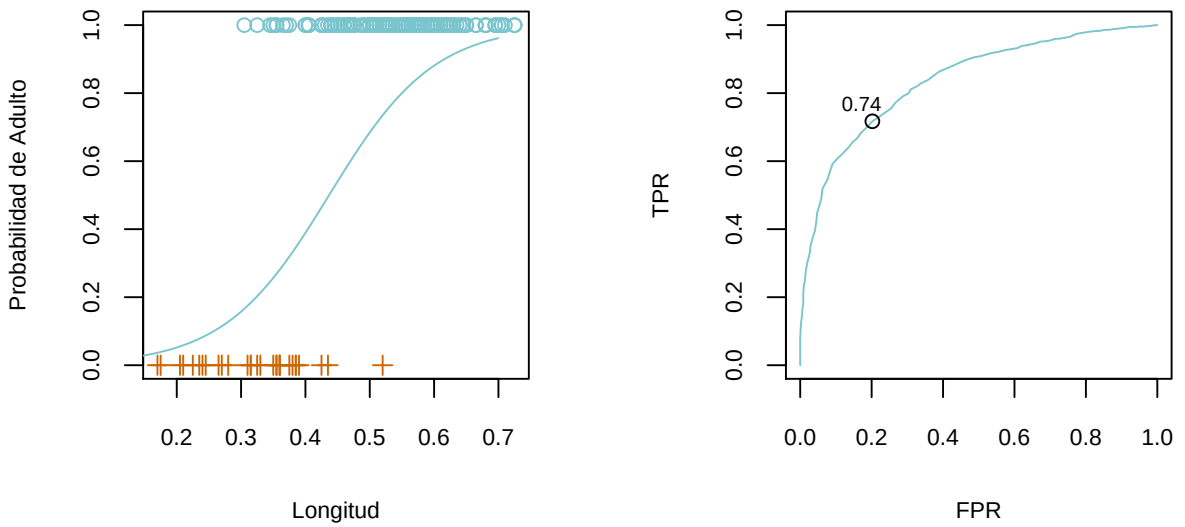
	No	Sí
No	232	116
Sí	71	417

Medida	Definición	Valor
Accuracy	$(TN + TP)/TOT$	0.78
Precision	$TP/(TP + FP)$	0.85
Recall	$TP/(TP + FN)$	0.78

Cuadro 2: **Izquierda:** Matriz de confusión comparando las predicciones del primer modelo con los resultados reales en el conjunto de entrenamiento `data.te`. **Derecha:** Medidas de desempeño del modelo de regresión logística utilizando la `longitud` como predictora sobre el conjunto de prueba `data.te`

.

Figura 2: **Izquierda:** Probabilidades estimadas de `adulto` por el modelo logístico. **Derecha:** Curva ROC para el clasificador logístico sobre `data.tr`. TPR es la fracción de casos clasificados correctamente como adultos. FPR la fracción de infantes clasificados incorrectamente como adultos. Se elige un umbral de decisión $p = 0{,}8$ cerca del extremo (0,1).



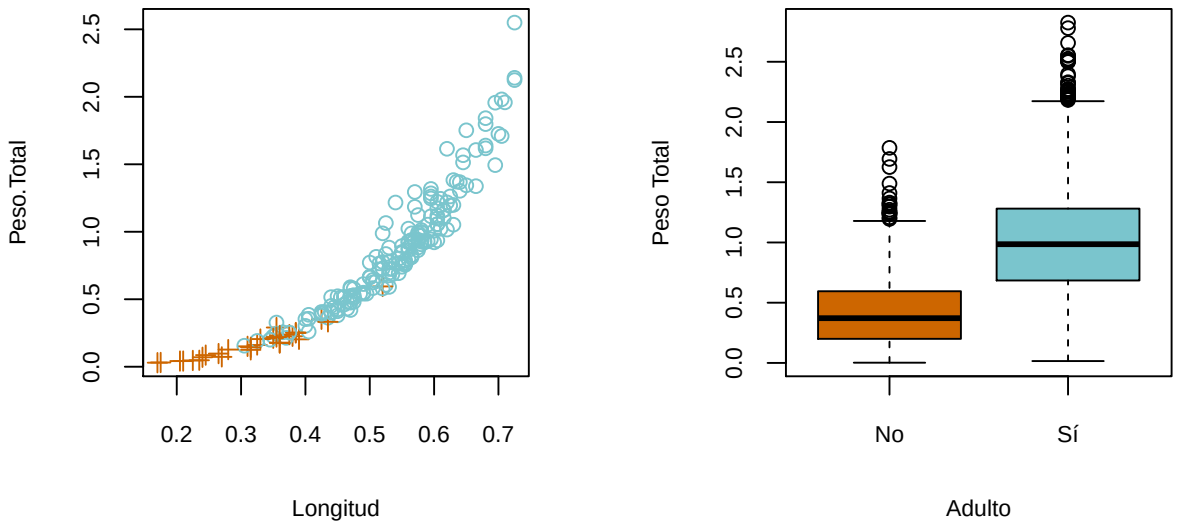


Figura 3: **Izquierda:** La longitud y peso total de doscientos especímenes. Los especímenes adultos se muestran en celeste, los infantes en naranja. **Derecha:** Boxplot del peso total como función de la variable binaria `adulto`.

2.2. Segundo Modelo Logístico - Clasificación por peso

En el panel izquierdo de la Figura 3 se muestran el `peso.total` y `longitud` de los primeros 200 casos del conjunto de datos `data.tr`. En el panel derecho de la Figura 3 se graficó el diagrama de cajas para la distribución de `peso.total` sobre la variable binaria `adulto`. Estos sugieren que el peso como predictora sea una buena forma de clasificar entre adultos e infantes de la forma

$$P(\texttt{adulto} = \texttt{Sí} | \texttt{peso.total}) = \frac{e^{\beta_0 + \beta_1 \texttt{peso.total}}}{1 + e^{\beta_0 + \beta_1 \texttt{peso.total}}} \tag{3}$$

En el Cuadro 3 se muestran los resultados de los estimadores de máxima verosimilitud para β_0 y β_1. Vemos que $\hat{\beta}_1 = 4{,}10$, indicando un aumento de la probabilidad de ser adulto ante incrementos positivos en el peso. El p-valor indica evidencia a favor de $\beta_1 \neq 0$, es decir, que la variable `peso.total` contribuye a explicar la variable `adulto`.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.96	0.10	-19.80	0.00
peso.total	4.10	0.15	27.06	0.00

Cuadro 3: Para data.tr, coeficientes estimados de regresión logística para la probabilidad de un espécimen sea adulto dado su peso total

En el gráfico izquierdo de la Figura 4 se muestra las probabilidades estimadas de $p(\texttt{adulto}|x) = 1$ según $x = \texttt{peso.total}$. En el gráfico derecho de la 4 se muestra la curva ROC del clasificador. Sobre ella eligió un valor de decisión $p = 0{,}71$ cercano al vertice (0,1). Utilizando este valor se evalua el rendimiento del clasificador sobre `data.te`. En el Cuadro 4 se resumen los resultados de la predicción, este modelo predice correctamente el 80 % de los casos, errando el 10 % restante.

	No	Sí
No	254	116
Sí	49	417

Medida	Definición	Valor
Accuracy	$(TN + TP)/TOT$	0.80
Precision	$TP/(TP + FP)$	0.89
Recall	$TP/(TP + FN)$	0.78

Cuadro 4: **Izquierda:** Matriz de confusión comparando las predicciones del segundo modelo con los resultados reales en el conjunto de entrenamiento `data.te`. **Derecha:** Medidas de desempeño del modelo de regresión logística utilizando la variable `peso.total` como predictora sobre el conjunto de prueba `data.te`

.

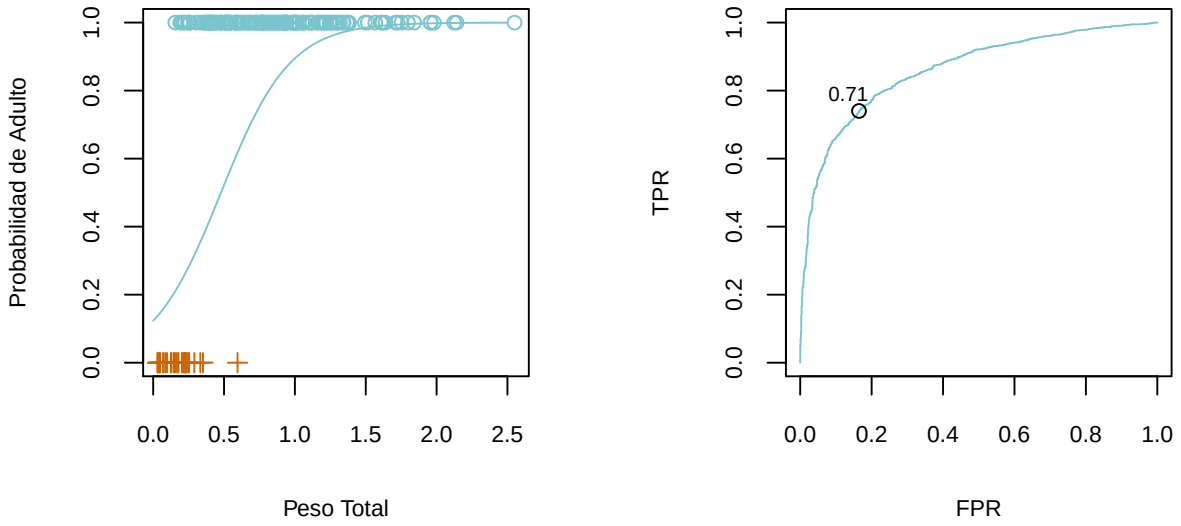


Figura 4: **Izquierda:** Probabilidades estimadas de `adulto` por el modelo logístico. **Derecha:** Curva ROC para el clasificador logístico sobre `data.tr`. TPR es la fracción de casos clasificados correctamente como adultos. FPR la fracción de infantes clasificados incorrectamente como adultos. Se elige un umbral de decisión $p = 0{,}71$ cerca del extremo (0,1).

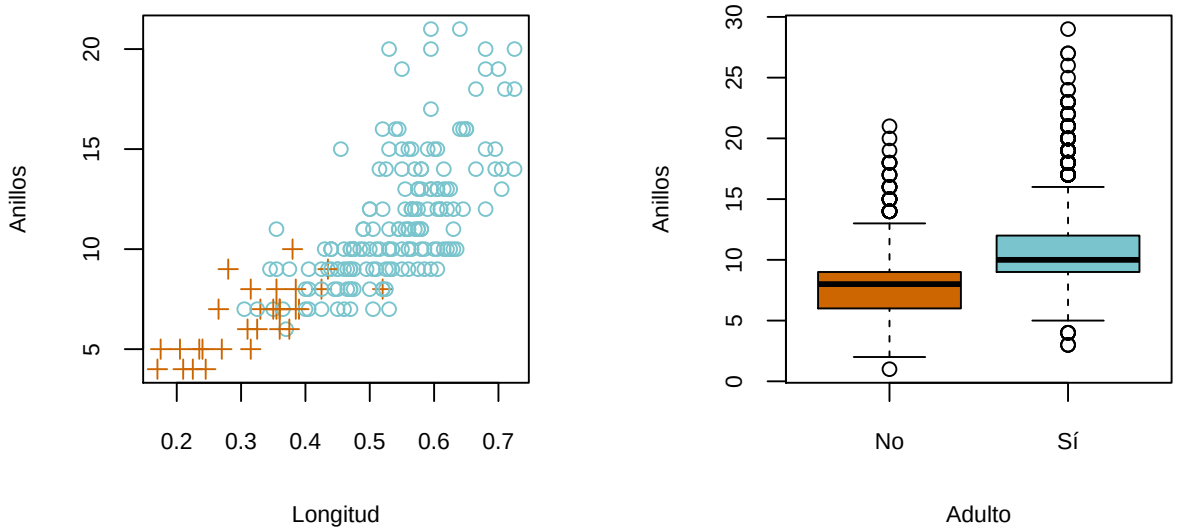
2.3. Tercer Modelo Logístico - Clasificación por anillos

De igual forma que en los casos anteriores realizamos otro modelo de regresión logística utilizando como predictora `anillos`. En la Figura 5 se muestran los datos de `data.tr` disponibles para entrenar. En el Cuadro 5 se muestran las estimaciones de coeficientes en $logit = \beta_0 + \beta_1$`anillos`. La variable `anillos` vuelve a ser significativa y con $\hat{\beta}_1 = 0{,}5$ indica variaciones positivas en *logit* ante variaciones positivas en `anillos`.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.80	0.19	-19.54	0.00
anillos	0.50	0.02	22.93	0.00

Cuadro 5: Coeficientes de regresión logística para la probabilidad de un espécimen sea adulto dado su cantidad de anillos estimados con los datos data.tr

Figura 5: **Izquierda:** La longitud y cantidad de anillos de doscientos especímenes en `data.tr`. Los especímenes adultos se muestran en celeste, los infantes en naranja. Se muestran solo 200 casos por claridad. **Derecha:** Boxplot de la cantidad de anillos como función de la variable binaria `adulto`.



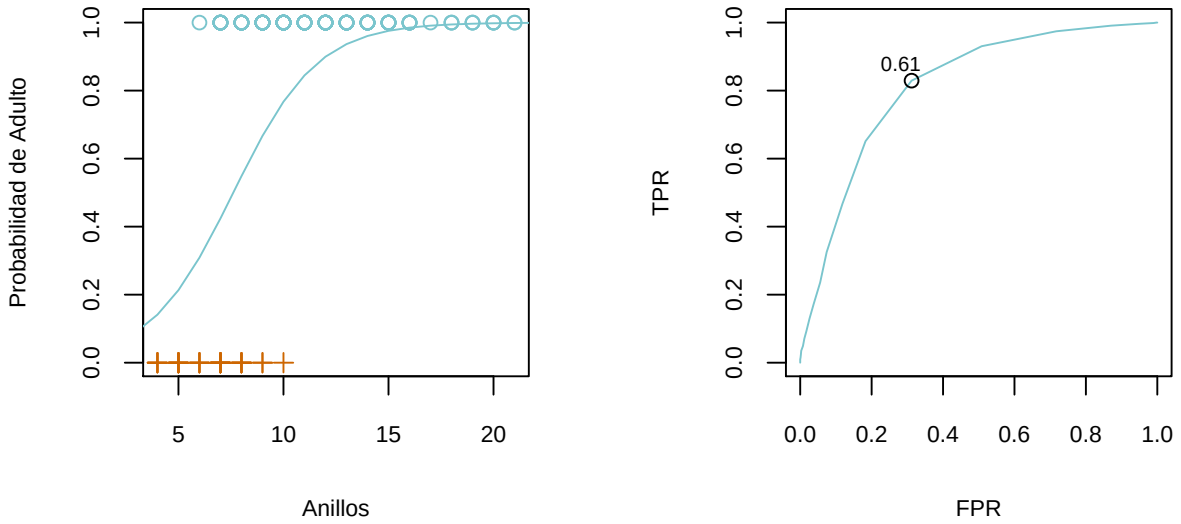


Figura 6: **Izquierda:** Probabilidades estimadas de `adulto` por el modelo logístico. **Derecha:** Curva ROC para el clasificador logístico sobre `data.tr`. TPR es la fracción de casos clasificados correctamente como adultos. FPR la fracción de infantes clasificados incorrectamente como adultos. Se elige un umbral de decisión $p = 0{,}61$ cerca del extremo (0,1).

En base a la curva ROC de la Figura 6 se elige el umbral de decisión $p = 0{,}61$. El el Cuadro 6 se resumen el rendimiento del clasificador sobre los datos de prueba `data.te`.

	No	Sí
No	207	92
Sí	96	441

Medida	Definición	Valor
Accuracy	$(TN + TP)/TOT$	0.78
Precision	$TP/(TP + FP)$	0.82
Recall	$TP/(TP + FN)$	0.83

Cuadro 6: **Izquierda:** Matriz de confusión comparando las predicciones del tercer modelo con los resultados reales en el conjunto de entrenamiento `data.te`. **Derecha:** Medidas de desempeño del modelo de regresión logística utilizando la variable `anillos` como predictora sobre el conjunto de prueba `data.te`

.

2.4. Cuarto Modelo Logístico - Clasificación por todas las medidas

El último modelo contempla todas las variables predictoras utilizadas, $logit = \beta_0 + \beta_1$`longitud` $+ \beta_2$`peso.total` $+ \beta_3$`anillos`. Los resultados del ajuste se muestran en el Cuadro 7. La probabilidad de umbral se elige según la curva ROC de la Figura 7 como $p = 0{,}63$. Evaluando el modelo en el conjunto de entrenamiento se obtienen los resultados del Cuadro 7.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.29	0.40	-0.73	0.47
longitud	-10.72	1.34	-7.98	0.00
peso.total	6.41	0.45	14.34	0.00
anillos	0.23	0.02	9.54	0.00

Cuadro 7: Para data.tr, coeficientes estimados de regresión logística para la probabilidad de un espécimen sea adulto dado su cantidad de anillos, peso total y longitud

Resulta interesante que en el Cuadro 7 el coeficiente de la variable longitud se vuelve negativo respecto al primer modelo. Mirando la matriz de correlación entre todas las predictoras vemos que `anillos` y `peso.total` tiene mucha correlación (0.89) lo que debe estar generando un efecto de colinealidad.

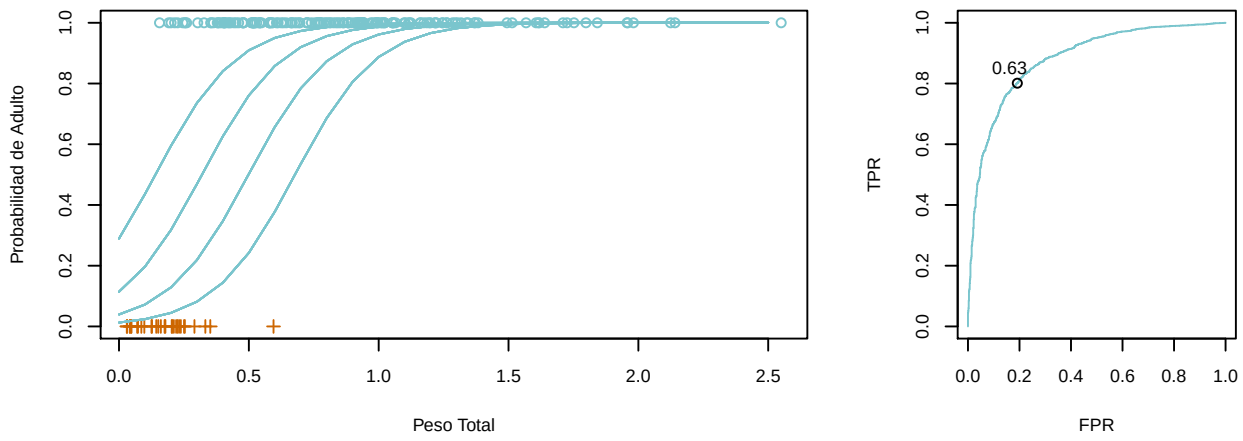


Figura 7: **Izquierda:** Probabilidades estimadas del modelo con todas la variables en función de `peso.total` con `longitud=0.4` y `anillos` $\in (1, 6, 11, 16)$. **Derecha:** Curva ROC para el modelo.

	No	Sí
No	237	81
Sí	66	452

Medida	Definición	Valor
Accuracy	$(TN+TP)/TOT$	0.82
Precision	$TP/(TP+FP)$	0.87
Recall	$TP/(TP+FN)$	0.85

Cuadro 8: **Izquierda:** Matriz de confusión comparando las predicciones del cuarto modelo con los resultados reales en el conjunto de entrenamiento `data.te`. **Derecha:** Medidas de desempeño del modelo de regresión logística utilizando todas las variables como predictoras sobre el conjunto de prueba `data.te`

.

3. Modelo LDA

De los modelos anteriores podemos estar seguros que todas las variables predictoras contribuyen a explicar el comportamiento de la variable `adulto`. Luego tiene sentido construir ahora un clasificador por *LDA* que utilice todas las predictoras. Entonces buscamos estimar las funciones discriminantes para las clases k, `adulto` e `infante`, y construir límites entre clases donde las funciones se igualan.

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - 0{,}5\mu_k^T\Sigma^{-1}\mu_k + log\pi_k \tag{4}$$

En el Cuadro 9 se muestran los parámetros estimados por frecuencias relativas sobre las observaciones del conjunto `data.tr`. La probabilidad a priori de la clase `adulto=Sí` estimada es $\pi_{Sí} \hat{=} 0{,}69$, es la probabilidad de que una muestra aleatoria provenga del grupo `adulto=Sí`, es decir, el 69 % del conjunto de entrenamiento es adulto. Para la clase `adulto=No` (Infante), $\pi_{\hat{No}} = 0{,}31$ (31 %). El Cuadro 9 también da las medias muestrales de cada predictora dentro de las clases, sugeriendo que las tres predictoras aumentan en casos de abulones adultos.

En la Figura 8 se muestra el umbral de decisión Bayesiano obtenido para el conjunto de entrenamiento dado por $(\mu_{Sí} + \hat{\mu}_{\hat{No}})/2$.

Para terminar y antes de evaluar elegimos un umbral de decisión para la probabilidad a posteriori $P($`adulto`$|$ (`longitud`, `peso.total`, `anillos`)). Para ello utilizamos nuevamente una curva ROC de la Figura 8 variando la probabilidad a posteriori de `adulto==Sí` y elegimos un valor de probabilidad 0,67 cercano al ideal $(0, 1)$. El Cuadro **??** muestra los resultados de aplicar con esta regla el modelo LDA sobre los datos de evaluación `data.te`. El modelo predice que de en la muestra hay 527 adultos, de los cuales 451 efectivamente lo son y 76 son infantes. Clasifica 309 como infantes, errando en 82 casos. Globalmente, la tasa de aciertos es 81 % y 18 % de fallas. Entre los infantes se clasificó incorrectamente el 26 %. Entre los adultos se clasificó incorrectamente el 14 %.

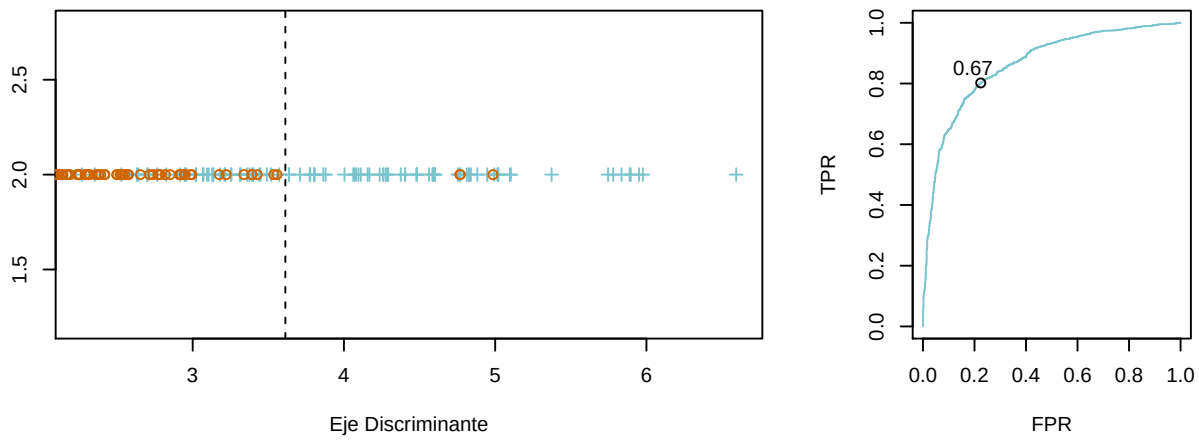


Figura 8: **Izquierda:** Umbral de decisión Bayesiano estimado para `data.tr`. Se grafican 100 especímenes del conjunto de training, en celeste los adultos y naranja los infantes. **Derecha:** Curva ROC sobre el conjunto de probabilidades a posteriori de `data.tr`. Para $\hat{p_k} = 0{,}67$ se minimiza la distancia a vértice (0,1).

	$\hat{\pi_k}$
No (Infante)	0.31
Sí (Adulto)	0.69

	$\hat{\mu_k}$		
	longitud	peso.total	anillos
No (Infante)	0.42	0.42	7.88
Sí (Adulto)	0.57	1.00	11.01

	Coef.
longitud	3.44
peso.total	1.15
anillos	0.11

Cuadro 9: Estimaciones de parámetros del modelo de LDA sobre el conjunto de entrenamiento `data.tr`. **Izquierda:** Estimaciones de probabilidades a *priori* de pertenencia a cada clase. **Centro:** Medias muestrales de cada predictora dentro de cada clase. **Derecha:** Coeficientes de discriminantes. Son los multiplicadores $3{,}44$`longitud` $+ 1{,}15$`peso.total` $+ 0{,}11$`anillos` que dan la mejor separación entre las dos clases

	No	Sí
No	227	82
Sí	76	451

Medida	Definición	Valor
Accuracy	$(TN + TP)/TOT$	0.81
Precision	$TP/(TP + FP)$	0.86
Recall	$TP/(TP + FN)$	0.85

Cuadro 10: Rendimiento del clasificador LDA sobre los datos `data.te`. **Izquierda:** Matriz de confusión para los 836 casos. **Derecha:** Medidas de desempeño del modelo de LDA.

Podemos intentar hacer al modelo de LDA un poco más flexible. Por intuición debe haber algún tipo de interacción entre `longitud` y `peso.total` debido a una densidad de la carne, lo que sugiere que puede ser apropiado incluir un término de interacción entre ellas `longitud:peso.total`. En el Cuadro 11 se muestran los resultados de evaluar este nuevo modelo sobre `data.te` con un umbral $p = 0{,}71$. Las predicciones mejoran ligeramente, intentar agregar más términos cruzados seguramente sea rebundante y resulte en *overfitting*

	No	Sí
No	232	74
Sí	71	459

Medida	Definición	Valor
Accuracy	$(TN + TP)/TOT$	0.83
Precision	$TP/(TP + FP)$	0.87
Recall	$TP/(TP + FN)$	0.86

Cuadro 11: Rendimiento del clasificador LDA sobre los datos `data.te`. **Izquierda:** Matriz de confusión para los 836 casos. **Derecha:** Medidas de desempeño del modelo de LDA.

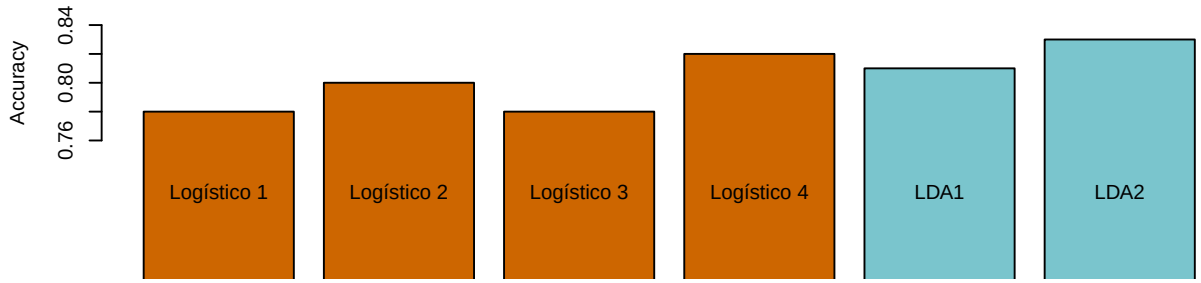


Figura 9: **Izquierda:** Comparación entre la medida *accuracy* de los distintos modelos sobre el conjunto de evaluación `data.te`.

4. Comparación de Modelos

En la Figura 9 se muestra la comparación del desempeño de los distintos modelos. Inicialmente se planteó el **criterio de maximizar la medida *accuracy* en la clasificación de los datos de** `data.te`. Con este criterio el mejor modelo es el LDA con una variable de interacción. Sin embargo todos los modelos tienen un buen desempeño, lo que sugiere que el verdadero límite de decisión para `adulto` que tratamos de estimar sea lineal como suponen los métodos de regresión logística y LDA, y que métodos más flexibles como QDA o KNN tengan resultados más pobres.

5. Bibliografía

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani - An Introduction to Statistical Learning with Applications in R (2015, Springer) - Capítulo IV
- Trevor Hastie, Robert Tibshirani, Jerome Friedman - The Elements of Statistical Learning, Data Mining, Inference, and Prediction. (2013, Springer) - Capítulo IV
- David G. Kleinbaum, Mitchel Klein - Logistic Regression A Self-learning Text (2002, Springer) - Capítulos I a III