DISPLASIA BRONCOPULMONAR EN PRETÉRMINOS: CLASIFICACIÓN BASADA EN VARIABLES CLÍNICAS A TRAVÉS DE MÉTODOS PARAMÉTRICOS Y NO PARAMÉTRICOS

Papalardo, Cecilia B. [1]; Chiapella Ferrari, Lilian [2]; Criado Blanco, Alexandra [2] ; Scavone Mauro, Cristina [3]

RESUMEN

La *Displasia Broncopulmonar (DBP)* es una enfermedad pulmonar crónica que ocurre con mayor frecuencia en Recién Nacidos Pretérminos (RNPT) que requirieron oxigenoterapia y ventilación mecánica. En el Servicio de Recién Nacidos del Centro Hospitalario Pereira Rossell (CHPR), que registra el mayor número de nacimientos en el Uruguay, muchos de los niños que nacen son prematuros con un peso al nacer menor o igual a 1500 g. Debido a su inmadurez, estos bebés tienen problemas sobre el sistema respiratorio, siendo su consecuencia mas frecuente el desarrollo de DBP. Con el fin de optimizar el tratamiento, control y seguimiento de estos pretérminos a comienzos del año 2008 se conformó un equipo interdisciplinario que desarrolló un proyecto financiado por la Comisión Sectorial de Investigación Científica (CSIC) denominado *"Evaluación saturométrica y polisomnográfica de prematuros portadores de displasia broncopulmonar"*. En el marco del proyecto se realizaron entre el 22 de abril del 2009 y el 31 de diciembre del 2011, oximetrías de pulso prolongadas a 210 RNPT con edad gestacional menor a 32 semanas y/o peso al nacer menor a 1500 g. Con la información recolectada en el proyecto se pretende construir, a partir de la aplicación de modelos paramétricos y no paramétricos como *Regresión Logística* y *Árboles de Regresión y Clasificación (CART)*, una regla que permita discriminar al grupo de los recién nacidos que presenta DBP de aquel que no presenta la enfermedad. Con esta regla y posibles modificaciones de la misma, se intenta asistir al personal del CHPR a clasificar a un RNPT dentro de un grupo relativamente homogéneo, a pocas horas del nacimiento, utilizando las variables clínicas disponibles y de esta forma permitirles planificar los niveles apropiados de cuidados del recién nacido, como por ejemplo, decidir si es necesario el aporte de oxígeno, la dosis y el tiempo que deberá recibirlo.

Palabras clave: *Displasia broncopulmonar, pretérminos, clasificación, CART.*

[1]Instituto de Estadística (IESTA) y Departamento de Métodos Cuantitativos - Area Matemática
[2]Escuela Universitaria de Tecnología Médica.
[3]Cátedra de Neuropediatría, Centro Hospitalario Pereira Rossell.

1. Introducción

La *Displasia Broncopulmonar (DBP)* es una enfermedad pulmonar crónica que ocurre con mayor frecuencia en Recién Nacidos Pretérminos (RNPT) con dificultad respiratoria aguda que requirieron oxigenoterapia y ventilación mecánica. Como consecuencia del avance de la Neonatología y de la mejoría en la prevención y el tratamiento de las complicaciones respiratorias en los RNPT, las manifestaciones clínicas de esta enfermedad han ido cambiado en el tiempo. En el año 2001, el consenso del National Institutes of Health (NIH) estableció que para el diagnóstico se requiere el antecedente de uso de oxígeno suplementario durante al menos 28 días.

El Servicio de Recién Nacidos del Centro Hospitalario Pereira Rossell (CHPR) registra el mayor número de nacimientos de Uruguay. Según cifras publicadas por la Fundación Álvarez Caldeyro Barcia (2011) en el año 2010 nacieron 7786 niños de los 47420 que lo hicieron en todo el país. El 12,1 % de los nacimientos ocurridos en el CHPR se producen antes de las 37 semanas de gestación, es decir, son bebes prematuros. Debido a la inmadurez, surgen problemas sobre el sistema respiratorio, siendo la consecuencia mas frecuente el desarrollo de DBP. Con el objetivo de optimizar el tratamiento, control y seguimiento de los niños que nacen pretérminos el CHPR conformó a comienzos del año 2008 un equipo interdisciplinario, el cual desarrolló un proyecto financiado por la Comisión Sectorial de Investigación Científica (CSIC) denominado *"Evaluación saturométrica y polisomnográfica de prematuros portadores de displasia broncopulmonar"*.

En el marco de este proyecto se obtuvo un conjunto de datos con información de las variables clínicas disponibles a pocas horas del nacimiento de 210 niños. El objetivo de este trabajo es explorar las características predictivas del problema e intentar construir una regla de clasificación que permita caracterizar al grupo de los recién nacidos que presenta DBP.

Los métodos estadísticos que se eligen tradicionalmente en estos casos son el Análisis Discriminante Lineal (ADL) y el análisis de Regresión Logística. Estos métodos hacen fuertes supuestos que en la mayoría de los casos no son válidos y presentan grandes desventajas, por ejemplo en el caso del ADL no es posible incluir variables del tipo categóricas que quizá pueden ser de gran utilidad para discriminar los grupos de interés. En los modelos de regresión se hacen fuertes supuestos de linealidad, siendo que en la mayoría de las aplicaciones en salud en las cuales el objetivo es caracterizar la ocurrencia de una respuesta es necesario lidiar con relaciones no lineales entre las variables exploratorias candidatas.

Para llegar a estos objetivos se utilizan métodos paramétricos y no paramétricos como *Regresión Logística* y *Árboles de Regresión y Clasificación (CART)*. Este último, pertenece al conjunto de técnicas de particionamiento recursivo las cuales no requieren explicitar la estructura del modelo. Las mismas se aplican para entender muchos de los problemas en ciencias biológicas, físicas y sociales, donde la relación entre las variables predictoras y la respuesta, es compleja.

Con esta regla y posibles modificaciones de la misma se intenta clasificar a un RNPT dentro de un grupo relativamente homogéneo a pocas horas del nacimiento sin tener que esperar 28 días de uso de oxígeno suplementario para tener que diagnosticar la enfermedad.

De esta forma, se busca asistir al personal del CHPR, brindándole mayor información al momento de tener que planificar los niveles apropiados de cuidados del recién nacido, como por ejemplo, decidir si es necesario el aporte de oxígeno, la dosis y el tiempo que deberá recibirlo.

2. Metodología

2.1. Regresión Logística

La regresión logística es un caso especial de los modelos lineales generalizados donde la función de enlace ("link" en inglés) que se utiliza es la función *logit*, de allí es que recibe su nombre. En este caso es de interés modelar una variable Y binaria que toma el valor 1 si ocurre determinado evento (éxito) y 0 si éste no ocurre (fracaso). Para cada individuo i se asume que la respuesta Y_i tiene distribución Bernoulli, o en forma equivalente, Binomial de parámetros 1 y p_i ($Y_i \sim \mathcal{B}(1, p_i)$), con función de probabilidad

$$P\{Y_i = y_i\} = p_i^{y_i}(1-p_i)^{1-y_i},\ y_i = 0, 1,\ i = 1, ..., n$$

donde los parámetros $p = (p_1, ..., p_n)'$ deben ser estimados con los datos. Para modelar estos datos, se intenta reducir los n parámetros en p a menos grados de libertad. La característica de la regresión logística es lograr este cometido modelando p_i con

$$p_i = p_i(\beta) = P(Y_i = 1|X = X_i) = \frac{e^{\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip}}}{1 + e^{\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip}}} \quad (1)$$

donde $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ es el nuevo vector de $(p + 1)$ parámetros a ser estimados y $X_i = (X_{i1}, ..., X_{ip})$ son los valores de las p covariables incluidas en el modelo para el $i - esimo$ individuo $(i = 1, ..., n)$.

En resumen: si p es la probabilidad de que ocurra el evento en cuestión

$$g : [0, 1] \to \mathcal{R} \text{ tal que } g(p) = \ln\left(\frac{p}{1-p}\right)$$

la función logística es la función inversa

$$g^{-1} : \mathcal{R} \to [0, 1] \text{ tal que } g^{-1}(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} = p$$

Así, si $p = P(Y = 1|_{X_1,X_2,...,X_p})$ el modelo de regresión logística está dado por

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Este último desarrollo ha sido extraído de las notas de Mesa (2006) donde también es posible encontrar el método de estimación de los parámetros del modelo, la interpretación en terminos de $\ln Oddsratio$ y el cálculo de la desvianza para comparar modelos y evaluar la calidad del ajuste. Además de este desarrollo teórico existen varias aplicaciones a distintos conjuntos de datos para ejemplificar el uso de los modelos de regresión logística.

2.2. Árboles de Regresión y Clasificación

Una de las aplicaciones de los métodos basados en árboles se encuentra cuando se intenta predecir la respuesta de una variable binaria Y basándose en la información que brindan p covariables $X_1, X_2, ..., X_p$.

Existen muchos algoritmos para la construcción de árboles de clasificación, pero en la mayoría se sigue una regla general: primero se particiona las observaciones utilizando una regla binaria en forma recursiva y segundo se ajusta un modelo constante en cada celda de la partición resultante. En el caso de variable respuesta binaria, el modelo a ajustar es $Y = 1$ ó $Y = 0$.

El primer paso de la regla se origina en el nodo raiz del árbol, en donde se encuentran todas las observaciones de la muestra de aprendizaje. El algoritmo selecciona una covariable X_j de las p disponibles y estima un punto de división que separe los valores de la respuesta Y_i en dos nodos hijos. Para una covariable X_j ordenada el punto de división es un número ξ que divide las observaciones en dos nodos. El primer nodo contiene todas las observaciones con $X_j \leq \xi$ y el segundo, contiene las observaciones que satisfacen que $X_j > \xi$. Para una covariable nominal X_j, los dos nodos se definen por el conjunto de niveles A, es decir si $X_j \in A$ ó $X_j \notin A$.

El objetivo del particionamiento recursivo es obtener nodos terminales lo más homogéneos posible en el sentido de que contengan sólo observaciones de uno sólo de los grupo ($Y = 1$ ó $Y = 0$). La homogeneidad completa de los nodos terminales rara vez se logra en el análisis de datos reales. De este modo, el objetivo de la partición es hacer la variable respuesta en los nodos terminales lo más homogénea posible.

Una medida cuantitativa del grado de homogeneidad del nodo es el concepto de impureza del nodo. La operación más sencilla es:

$$\frac{\text{Número de observaciones con } \ Y = 1 \ \text{ en un nodo}}{\text{Número total de observaciones en el nodo}}$$

Cuanto más cerca esta relación de 0 ó 1, más homogéneo es el nodo. Una variación de esta medida de impureza es el índice de Gini.

Una vez que la división ξ o A es estimada para alguna covariable X_j, se aplica el mismo procedimiento en cada uno de los dos nodos hijos obtenidos. La recursión termina cuando se cumple algún criterio de parada exigido. Pero decidir sobre este punto, no es trivial. De hecho, los árboles con muchas hojas pueden sufrir un sobreajuste y árboles pequeños pueden perder aspectos importantes del problema. Una estrategia a utilizar es dejar crecer el árbol usando un criterio de parada trivial, como el número de observaciones en una hoja, y luego podar las ramas que no son necesarias.

En general, la mayoría de los algoritmos disponibles difieren respecto a tres puntos: (1) cómo la covariable es seleccionada en cada paso, (2) como el punto de división es estimado y (3) qué criterio de parada es aplicado. Uno de los algoritmos más populares fue descrito por Breiman et al. (1984) y está disponible en R en la biblioteca *rpart*. Este algoritmo primero examina todas los posibles divisiones para todas las covariables y elige la división que permita que los dos nodos otenidos sean más "puros" que el actual, con respecto a

los valores de la variable respuesta Y. Hay muchas medidas de impureza disponibles, para problemas de regresión con respuesta nominal el criterio por defecto en *rpart* es el de *Gini*. Lo desarrollado en esta sección puede ser encontrado con mayor detalle en Zhang y Singer (2010) y Everitt y Ohothorn (2010).

3. Resultados

Se trabaja con un conjunto de datos de 210 niños que nacieron con peso menor a 1500 g y/o edad gestacional menor a 32 semanas y que recibieron una oximetría de pulso prolongada entre el 22 de abril del 2009 y el 31 de diciembre del 2011. La variable respuesta binaria es el padecimiento de broncodisplasia (Si $= 1$, No $= 0$). Los predictores o variables explicativas que se utilizan se listan en el siguiente cuadro:

Variable	**Etiqueta**	**Tipo**	**Valores**
Lugar del que proviene	`lugar`	Categórica	Montevideo Interior
Edad de la madre al momento del embarazo	`edadmademb`	Cuantitativa	13 - 46
Número de gestaciones	`ngestas`	Cuantitativa	1 - 14
Edad gestacional (semanas)	`eg`	Cuantitativa	24 - 34
Cantidad de controles según edad gestacional	`conteg`	Cuantitativa	0 - 0.4
Genero	`genero`	Categórica	Femenino Masculino
Peso al nacer (gramos)	`peso`	Cuantitativa	615 - 1840
Tipo de parto	`parto`	Categórica	Vaginal Cesárea
Talla (cm)	`talla`	Cuantitativa	28 - 47
Perímetro craneano (cm)	`pc`	Cuantitativa	20.5 - 35.3
Recibió antibióticos	`atb`	Categórica	No - Si
Recibió corticoides prenatales	`corticoideprenat`	Categórica	No - Si
Recibió surfactante	`surfactante`	Categórica	No - Si
Recibió aminofilina	`aminofilina`	Categórica	No - Si

Cuadro 1: Descripción de cada una de las variables del conjunto de datos.

De los 210 niños, 64 recibieron el diagnóstico de broncodisplasia.

En base a esta información se tratará de dar respuesta a los siguientes interrogantes:

- ¿Es posible caracterizar a los niños broncodisplásicos con un margen de error aceptable, a partir de estas variables?
- En ese caso, ¿cuáles de ellas son las más importantes?

Previamente se explora la relación existente entre las variables predictivas a través de una matriz de dispersión, que también presenta las correlaciones muestrales correspondientes. Luego, a través de sucesivos diagramas de caja se visualiza el comportamiento de cada variable predictiva según los dos grupos de interés.

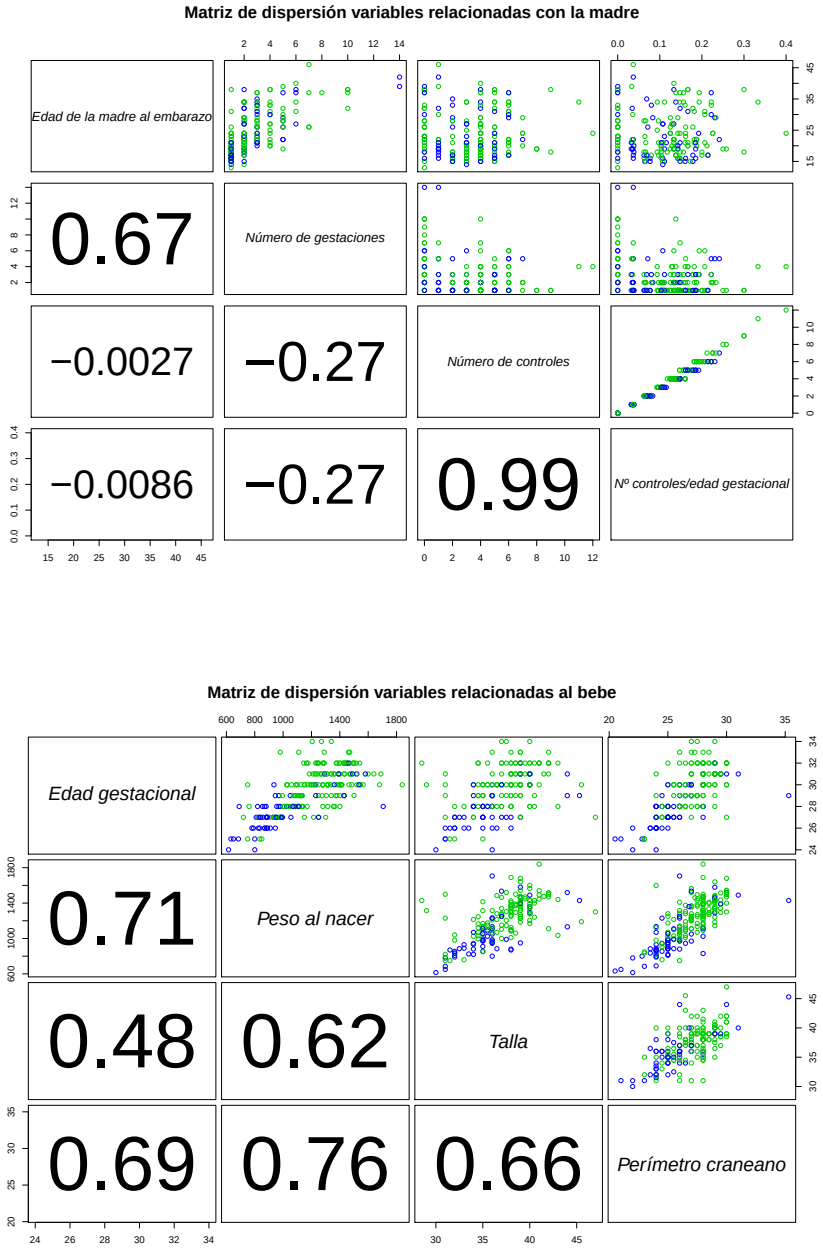

Figura 1: Matriz de dispersión y correlaciones muestrales para las variables cuantitativas.

Las parejas de variables peso al nacer y perímetro craneano, peso al nacer y edad gestacional son las que presentan correlaciones más altas de 0.76 y 0.71 respectivamente. En general las variables que tienen que ver con el tamaño, la forma y el peso del niño están correlacionadas entre sí, es decir, niños con edad gestacional mayor, son los que presentan mayor peso al nacer, perímetro craneano y talla mayor.

Entre las variables que corresponden a la madre, las dos más correlacionadas son la edad de la madre al momento del embarazo y el número de gestaciones previas (0.67), lo cual parece tener sentido. La alta correlación observada entre la variable número de controles y la cantidad de controles según la edad gestacional (0.99), se debe a la forma de construir esta última a partir del cociente entre la primera y la edad gestacional. El número de controles realizados durante el embarazo no fue listado entre las variables predictivas, debido a que esta variable no fue utilizada en los ajustes de los modelos.

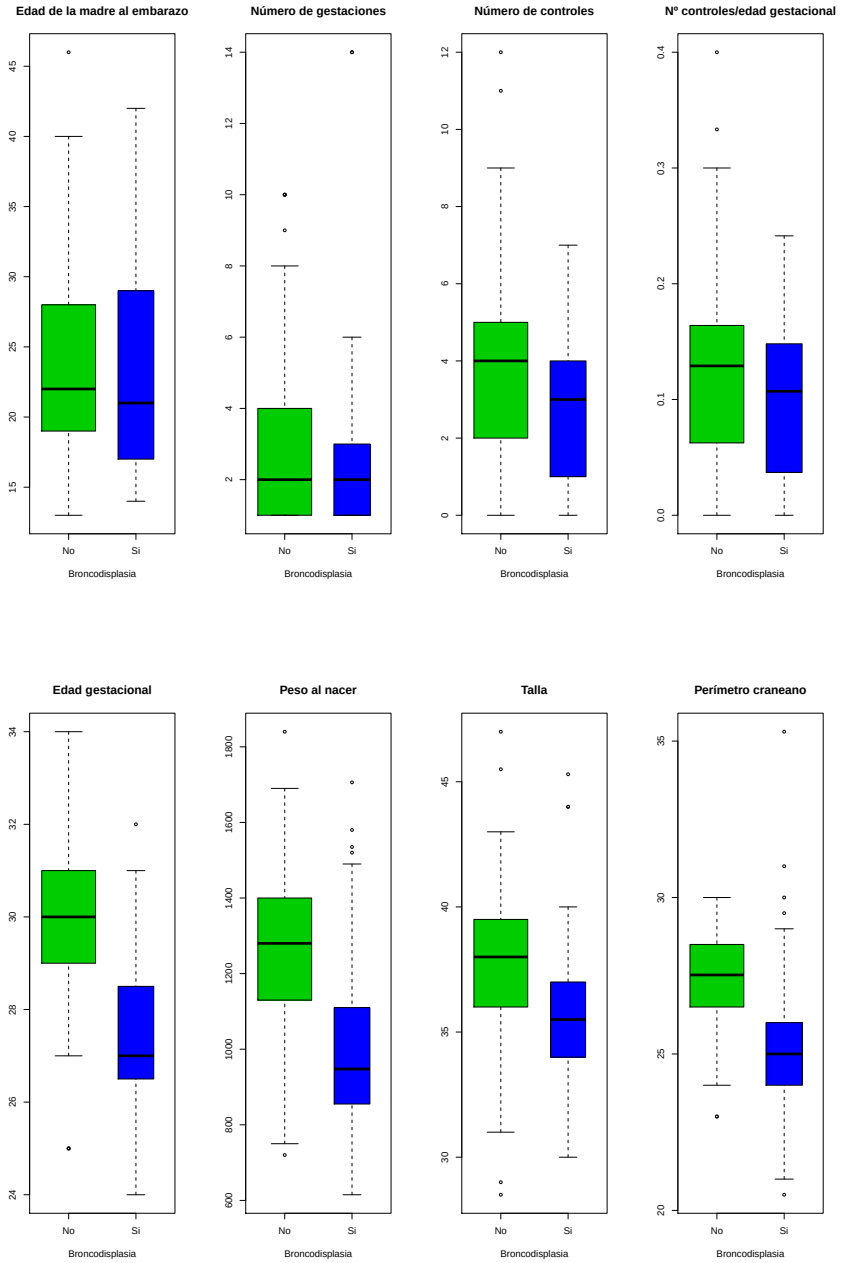

Figura 2: Diagramas de cajas para las variables cuantitativas según la variable de interés.

Las variables que parecen diferenciar a los dos grupos, son las que están vinculadas directamente con el niño (edad gestacional, peso al nacer, etc.)

Regresión Logística

Se quiere predecir la probabilidad de padecer broncodisplasia en función de las variables clínicas disponibles al momento del parto. Para lograr este objetivo se ajusta el modelo de regresión logística que contiene los efectos principales utilizando la función *glm* del software libre R versión 2.13.1. Para comparar los distintos modelos a través del criterio AIC (Criterio de Información de Akaike), se utiliza la función *bestglm*.

Debido a la falta de información en alguna de las variables predictoras, 49 de los 210 casos no son utilizados en el análisis. Además se realiza una separación de los datos en muestra de entrenamiento o aprendizaje (120 casos) y muestra de prueba o test (41 casos). La primera muestra es utilizada para construir el modelo, la otra para ponerlo a prueba. Esta es la estrategia más simple que permite crear una validación artificial del estudio, tiene como costo reducir el tamaño de muestra para estimar el modelo, pero asegura una mejor aproximación a los errores de predicción.

El modelo seleccionado es

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1 \texttt{eg} + \beta_2 \texttt{ngestas} + \beta_3 \texttt{conteg} + \beta_4 \texttt{surfactantesi}$$

Se realiza la comparación de los dos modelos anidados, el que contiene solo el intercepto y el que contiene además las 4 variables elegidas, usando la función *anova*. Los resultados obtenidos

```
Analysis of Deviance Table

Model 1: dbp ~ 1
Model 2: dbp ~ ngestas + conteg + eg + surfactante
       Resid.  Df Resid.Dev  Df Deviance
1       119     149.84
2       115     102.76       4    47.08
```

El valor obtenido de la Desvianza ("Deviance" en inglés) se compara con el valor de tabla de un χ^2 con 4 grados de libertad (9.49), resultando significativo al 5 % de significación. Esto permite testear la posibilidad de que $\beta_i \neq 0$ para algún i con $i = 1, ..., 4$. Es decir al menos uno de los parámetros (o quizá todos) es distinto de cero.

En el Cuadro 2 se presentan los resultados que devuelve la función *bestglm* con la estimación de los coeficientes de los predictores seleccionados, seguido por los errores estándar de los coeficientes estimados, el valor del estadístico z y el valor p-asociado de la prueba z para testear si cada coeficiente es o no distinto de cero.

	Estimate	**Std. Error**	**z value**	**Pr**$(>\|\mathbf{z}\|)$
(Intercept)	19.8989891	4.7642565	4.176725	2.96E-05
`eg`	-0.7544959	0.1593151	-4.73587	2.18E-06
`ngestas`	-0.2424641	0.1194999	-2.028991	4.25E-02
`conteg`	-6.404381	3.8226786	-1.675365	9.39E-02
`surfactantesi`	2.5313674	1.4596835	1.734189	8.29E-02

Cuadro 2: Coeficientes para el mejor modelo según criterio AIC.

Según el modelo de regresión logística las variables edad gestacional, número de gestaciones, número de controles en relación con la edad gestacional y el uso de surfactante, son las variables a tener en cuenta al momento de predecir la probabilidad de padecer broncodisplasia. Para este modelo se calcula el error de predicción sobre las observaciones en la muestra de entrenamiento (Cuadro 3) y la muestra de test (Cuadro 4).

		Predicho		
		No	Si	Total
Real	No	71	11	82
	Si	16	22	38
	Total	87	33	120

Cuadro 3: Clasificación de la muestra de entrenamiento en regresión logística.

$$\text{Error de clasificación} = \frac{11+16}{120} \times 100 = 22.5\,\%$$

		Predicho		
		No	Si	Total
Real	No	27	1	28
	Si	6	7	13
	Total	33	8	41

Cuadro 4: Clasificación de la muestra de test en regresión logística.

$$\text{Error de clasificación} = \frac{1+6}{41} \times 100 = 17.07\,\%$$

Al final de la sección se presenta un cuadro que compara los porcentajes de acierto de los modelos utilizados, tanto a nivel global como para cada una de las categorías de interés.

Los datos faltantes pueden provocar serias pérdidas de información. En el análisis siguiente se utilizan métodos basados en árboles, que hacen un manejo eficiente de los datos faltantes creando una categoría distinta para estos valores.

Árboles de Regresión y Clasificación

Se construye un árbol utilizando la biblioteca *rpart* del software libre R versión 2.13.1. En el mismo se especifica que se siga particionando cada nodo sólo si se cuenta con por lo menos 5 observaciones, además el índice de gini es utilizado como medida de impureza. Previamente, se realiza una separación de los datos en muestra de entrenamiento (140 casos) y muestra de test (70 casos). La primer muestra es utilizada para construir el árbol, la otra para ponerlo a prueba.

La representación grafica del árbol fue obtenida a través de la biblioteca *partykit* de R y se presenta en la Figura 3.

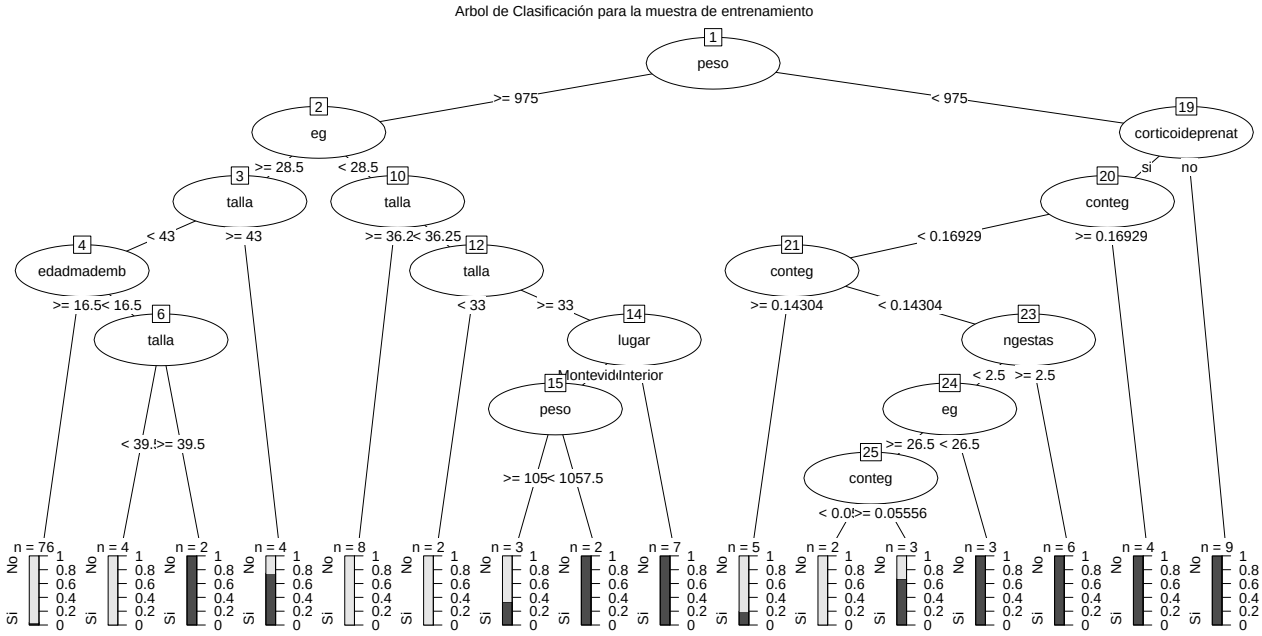

Figura 3: Árbol de Clasificación construido con los 140 niños de la muestra de entrenamiento.

La primer variable utilizada por el árbol es el peso al nacer, por lo cual es la variable que mejor divide a los grupos según este método.

Para caracterizar los nodos terminales debe seguirse el camino de condiciones establecidas sobre los datos. A modo de ejemplo, se interpreta el primer nodo a la derecha del árbol. El mismo contiene 9 niños, todos broncodisplásicos, que según las particiones que se realizaron se caracterizan por nacer con bajo peso (menos de 997.5 g) y no haber recibido corticoides prenatales. Estas condiciones son razonables en el contexto del problema.

En el Cuadro 5 se presenta el resultado global de utilizar el árbol para clasificar a los niños en la muestra de entrenamiento. Es importante tener en cuenta que si se utiliza el árbol construido para clasificar a otro conjunto de niños recién nacidos, en broncodisplásicos o no, el error de clasificación puede ser mayor. Tratando de hacer una mejor aproximación se calcula el error de clasificación sobre la muestra de test (Cuadro 6).

		Predicho		
		No	Si	Total
Real	No	95	2	97
	Si	5	38	43
	Total	100	40	140

Cuadro 5: Clasificación en la muestra de entrenamiento en CART.

$$\text{Error de clasificación} = \frac{2+5}{140} \times 100 = 5.00\,\%$$

		Predicho		
		No	Si	Total
Real	No	44	5	49
	Si	7	14	21
	Total	51	19	70

Cuadro 6: Clasificación en la muestra de test en CART.

$$\text{Error de clasificación} = \frac{7+5}{70} \times 100 = 17.14\,\%$$

Comparación de los porcentajes de aciertos de los modelos utilizados

A modo de resumen se presenta en el Cuadro 7 los porcentajes de acierto del modelo de regresión logística ajustado y del árbol de clasificación construido, tanto para la muestra de entrenamiento como para la muestra de test.

		Acierto		
Modelo	Muestra	Global	No	Si
Regresión	Ent (120)	77.5	86.6	57.9
Logística	Test (41)	82.9	96.4	53.8
Árbol de	Ent (140)	95	97.9	88.4
Clasificación	Test (70)	82.9	89.8	66.7

Cuadro 7: Porcentajes de acierto.

4. Consideraciones finales

Una desventaja de los árboles de clasificación utilizados consiste en la inestabilidad de la construcción debido a su sensibilidad a pequeños cambios en el conjunto de datos analizados. Por ejemplo, al agregar algunos casos nuevos se producen modificaciones mayores en el árbol resultante.

Se deja planteada la necesidad de seguir trabajando en la construcción de un árbol que permita mejorar la clasificación de futuros casos, con un menor margen de error, por ejemplo, a través de un procedimiento denominado proceso de poda. Por otro lado, existen extensiones de las técnicas basadas en árboles que pueden ayudar con este problema, como Bagging (Bootstrap Aggregating) y Bosques Aleatorios (Random Forest), que plantean la construcción de varios árboles y su combinación para la predicción o clasificación.

5. Bibliografía

Everitt, B. y Ohothorn, T. (2010). *A Handbook of Statistical Analysis using R*. Chapman & Hall/CRC, Boca Raton, 2ª ed..

Fundación Álvarez Caldeyro Barcia (2011). Nacer en tiempo. `http://www.fundacionalvarezcaldeyrobarcia.org.uy/`

Mesa, Andrea (2006). Modelos Lineales Generalizados. *Informe técnico*, Laboratorio de Probabilidad y Estadística. Facultad de Ingeniería. UdelaR.

Zhang, H. y Singer, B. H. (2010). *Recursive Partitioning and Applications*. Springer, New York, 2ª ed..