

Universidad Nacional De La Matanza

Especialización Ciencia de Datos

Fundamentos para la Ciencia de Datos

Profesor: Silvia N. Pérez

Modelado de la Miopía Infantil a partir de Variables Biométricas y de Estilo de Vida

Análisis Exploratorio, Regresión Logística y Naive Bayes
aplicado a un dataset clínico de niños

Integrantes Grupo 10:

Fica Millán, Yesica	DNI 27.624.956
Miranda Charca, Florencia	DNI 41.398.768
Petraroia, Franco	DNI 27.161.862

Tabla de contenido

1. Introducción	5
2. Análisis Exploratorio de Datos (EDA).....	6
2.1. Importación y descripción inicial del dataset.....	6
2.2. Revisión de Datos Faltantes y Duplicados.....	7
2.3. Detección de Valores Atípicos.....	7
2.4. Recodificación de Variables Categóricas.....	9
2.5. Clasificación de variables por tipo y rol.....	9
2.6. Resumen.....	10
3. Preparación del Dataset	10
3.1. Generación de la Partición Estratificada	10
3.2 Distribución de clases por conjunto	11
3.3 Visualización	11
4. Modelado con Regresión Logística.....	12
4.1. Modelo 1: Con Todas las Variables	12
4.2. Modelo 2: Selección Automática (Stepwise).....	13
4.3. Modelo 3: Modelo Propuesto por el Grupo.....	13
4.4. Comparación y selección del modelo:.....	14
5. Evaluación del Modelo Elegido	15
5.1. Significancia de Variables	15
5.2. Test de Bondad de Ajuste.....	16
5.3. Interpretación de Coeficientes.....	16
5.4. Análisis de Multicolinealidad.....	17
5.5. Influencia de Casos.....	17

5.6. Síntesis del Modelo Seleccionado	18
6. Modelado con Naive Bayes	19
6.1. Ajuste del modelo	19
6.2. Validación del modelo en el conjunto test.....	20
6.3. Curva ROC y AUC	21
6.4. Síntesis interpretativa	22
7. Evaluación y Comparación de Modelos en el Conjunto Test	22
7.1. Curva ROC y AUC	22
7.2. Tabla de Clasificación y Métricas	24
7.4. Discusión de Resultados	24
8. Elección del Modelo Según el Objetivo.....	25
8.1. Elección del Punto de Corte Óptimo	25
8.2. Comparación del Modelo 2 con distintos umbrales y Naive Bayes	26
8.3. Elección del Modelo para Detectar la Mayor Cantidad de Niños con Miopía	28
8.4. Elección Modelo para Minimizar Errores Globales	28
9. Conclusiones Finales.....	29
10. Anexos	30
11. Referencias	30

Índice de Gráficos

Ilustración 1 - Boxplot de variables numéricas	8
Ilustración 2 - Proporción de miopía en conjunto de entrenamiento y validación	11
Ilustración 3 - Distancia de Cook por observación	18
Ilustración 4 - Curva ROC - Naive Bayes	22
Ilustración 5 - Curvas ROC - comparación de Modelos	23
Ilustración 6: Métricas según el umbral de corte.....	27
Ilustración 7: Comparación de Desempeño de Modelos de Clasificación mediante Métricas Normalizadas.....	28

1. Introducción

Este trabajo propone un análisis estadístico aplicado al diagnóstico de miopía infantil, utilizando datos clínicos provenientes de un estudio realizado en niños. El análisis se desarrollará a partir del dataset `chicos25.xlsx`, que contiene información relevante para identificar factores asociados a la presencia de esta condición visual.

El diagnóstico temprano de la miopía en niños permite intervenir a tiempo con tratamientos preventivos. Por ello, un modelo predictivo preciso podría ser de gran utilidad para priorizar estudios clínicos.

Objetivo del Trabajo

El objetivo principal es construir y comparar diferentes modelos estadísticos que permitan predecir la presencia de miopía a partir de las variables disponibles en el conjunto de datos. A través de este enfoque, se busca identificar los factores más asociados con esta condición y evaluar el desempeño predictivo de cada modelo.

Descripción General del Dataset

El conjunto de datos utilizado contiene 618 observaciones y 9 variables, y fue proporcionado en formato de tabla bajo el nombre `chicos25.xlsx`.

Este dataset recoge información clínica y conductual de niños que participaron en un estudio sobre miopía, con el objetivo de identificar posibles factores asociados a esta condición ocular.

A continuación se presenta un resumen de las variables incluidas:

Nº	VARIABLE	TIPO	EJEMPLO	DESCRIPCIÓN BREVE
1	id	numérica	1, 2, 3	Identificador único del individuo
2	mio	categorica	Sí, No, No	Diagnóstico de miopía
3	spheq	numérica	-0.052, 0.608, 1.179	Medidas Optométricas
4	Al	numérica	21.89, 22.38, 22.49	Medidas Optométricas
5	acd	numérica	3.69, 3.70, 3.46	Medidas Optométricas
6	vcd	numérica	14.7, 15.29, 15.52	Medidas Optométricas
7	sporthr	numérica	40, 4, 14	Horas mensuales dedicadas al deporte
8	tvhr	numérica	10, 7, 10	Horas mensuales dedicadas a la televisión
9	Pa	categorica	Sí, Sí, No	El padre tiene miopía (Sí / No)

Esta información será utilizada para realizar el análisis exploratorio de datos (EDA) con el fin de comprender la distribución de las variables, identificar posibles asociaciones relevantes y preparar el terreno para el ajuste de modelos de regresión logística, orientados a predecir la presencia de miopía.

2. Análisis Exploratorio de Datos (EDA)

2.1. Importación y descripción inicial del dataset

Para comenzar con el análisis, se procedió a cargar el conjunto de datos `chicos25.xlsx` el cual contiene la información de 618 individuos, estructurada en 9 variables.

Para facilitar la interpretación de los resultados y mejorar la legibilidad durante el análisis, se renombraron las variables originales del dataset de la siguiente manera:

NOMBRE ORIGINAL	NUEVO NOMBRE
mio	miopia
spheq	med_spheq
al	med_al
acd	med_acd
vcd	med_vcd
sporthr	hs_mes_deporte
tvhr	hs_mes_tv
pa	padre_miopia
id	id (sin cambios)

Posteriormente, se inspeccionó la estructura del dataset con la función `str()`. Se identificaron dos variables categóricas (`miopia` y `padre_miopia`) originalmente codificadas como `character`, que fueron convertidas a `factor`. Las variables restantes fueron correctamente reconocidas como numéricas.

Se presentó una vista preliminar de las primeras observaciones y se generaron resúmenes estadísticos y descriptivos por variable, utilizando funciones como `summary()`, `skim()` y `dfSummary()`.

VARIABLE	TIPO INICIAL	OBSERVACIÓN
MIOPIA	character	Convertida a <code>factor</code>
PADRE_MIOPIA	character	Convertida a <code>factor</code>
RESTO DE VARIABLES	numeric	Se mantienen como numéricas

Estas transformaciones permiten el correcto tratamiento estadístico en modelos de clasificación como regresión logística o Naive Bayes, los cuales requieren que las variables categóricas estén codificadas como `factor`.

A continuación, se presenta un resumen descriptivo detallado de las variables incluidas en el conjunto de datos, excluyendo el identificador individual (`id`). Esta tabla, ofrece una visión general de la estructura de las variables, incluyendo el tipo de dato, cantidad de valores únicos, proporción de valores faltantes, medidas de tendencia central y dispersión, así como una visualización de la distribución de cada variable. Esta información resulta fundamental para comprender la naturaleza del conjunto de datos antes de avanzar con el análisis estadístico.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	miopia [factor]	1. No 2. Si	537 (86.9%) 81 (13.1%)		618 (100.0%)	0 (0.0%)
2	med_speq [numeric]	Mean (sd) : 0.8 (0.6) min ≤ med ≤ max: -0.7 ≤ 0.7 ≤ 4.4 IQR (CV) : 0.6 (0.8)	511 distinct values		618 (100.0%)	0 (0.0%)
3	med_al [numeric]	Mean (sd) : 22.5 (0.7) min ≤ med ≤ max: 19.9 ≤ 22.5 ≤ 24.6 IQR (CV) : 0.9 (0)	254 distinct values		618 (100.0%)	0 (0.0%)
4	med_acd [numeric]	Mean (sd) : 3.6 (0.2) min ≤ med ≤ max: 2.8 ≤ 3.6 ≤ 4.2 IQR (CV) : 0.3 (0.1)	206 distinct values		618 (100.0%)	0 (0.0%)
5	med_vcd [numeric]	Mean (sd) : 15.4 (0.7) min ≤ med ≤ max: 13.4 ≤ 15.4 ≤ 17.3 IQR (CV) : 0.9 (0)	226 distinct values		618 (100.0%)	0 (0.0%)
6	hs_mes_deporte [numeric]	Mean (sd) : 12.5 (19.3) min ≤ med ≤ max: 0 ≤ 10 ≤ 450 IQR (CV) : 10 (1.5)	39 distinct values		618 (100.0%)	0 (0.0%)
7	hs_mes_tv [numeric]	Mean (sd) : 8.9 (5.7) min ≤ med ≤ max: 0 ≤ 8 ≤ 31 IQR (CV) : 7.8 (0.6)	28 distinct values		618 (100.0%)	0 (0.0%)
8	padre_miopia [factor]	1. No 2. Si	310 (50.2%) 308 (49.8%)		618 (100.0%)	0 (0.0%)

2.2. Revisión de Datos Faltantes y Duplicados

Para evitar errores en las estimaciones de los modelos a desarrollar se procedió a controlar si existen valores faltantes (NA) o duplicados.

- En el análisis realizado no se encontraron este tipo de valores.

2.3. Detección de Valores Atípicos

Para la detección de valores atípicos se construyeron gráficos de tipo boxplot para cada una de las variables numéricas. Este enfoque permitió identificar posibles registros extremos que pudieran corresponder a errores de medición o carga.

Durante la exploración de la variable `hs_mes_deporte` (horas mensuales dedicadas a actividad física), se detectó un valor atípico extremo de 450 horas, lo que representa más de 15 horas diarias de ejercicio. Este valor resulta clínicamente inverosímil en el contexto de un estudio con niños pequeños, por lo que se considera un error de carga o medición.

- El tratamiento específico para este caso se expone en la etapa de limpieza de datos.

En cuanto al resto de las variables, si bien se observaron algunos valores extremos, estos fueron considerados plausibles dentro del marco clínico del estudio, ya que podrían corresponder a características propias de niños con diagnóstico de miopía. Por lo tanto, no se consideran errores, sino parte de la variabilidad esperada en la población bajo análisis.

Con el objetivo de explorar la distribución de las variables numéricas presentes en el conjunto de datos, se construyeron diagramas de caja para cada una de ellas. Esta visualización permite identificar rápidamente la dispersión, la presencia de asimetrías y posibles valores atípicos.

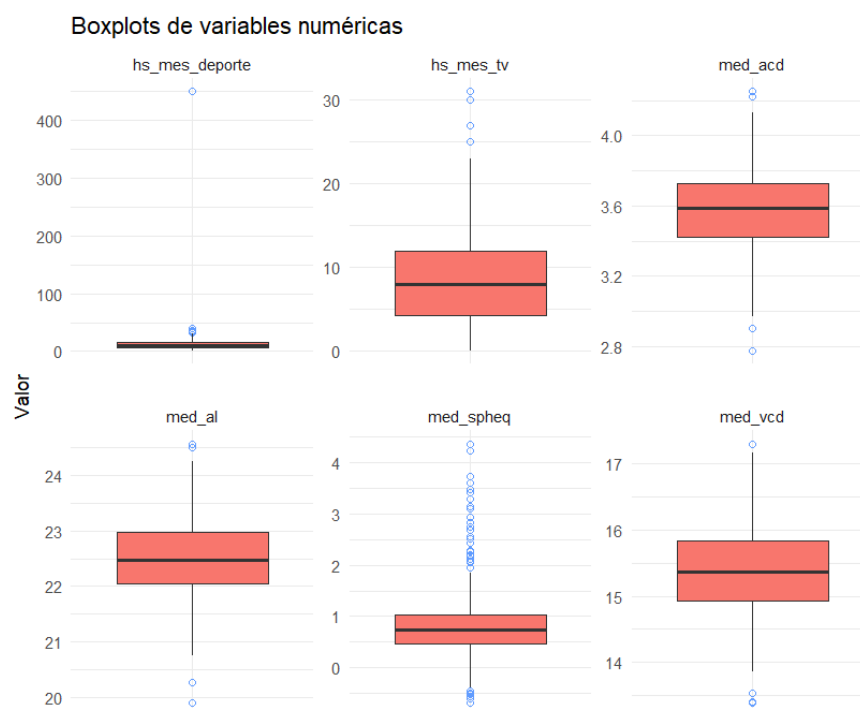


Ilustración 1 - Boxplot de variables numéricas

Como parte del análisis exploratorio, se identificó la observación con la mayor cantidad registrada de horas mensuales dedicadas a actividad física. Esta búsqueda permite detectar valores potencialmente atípicos que podrían influir en el ajuste de los modelos estadísticos. A continuación, se detalla la observación correspondiente al máximo valor en la variable `hs_mes_deporte`.

Observación con el máximo valor de horas mensuales de actividad física

	<code>id</code>	<code>miopia</code>	<code>med_spheq</code>	<code>med_al</code>	<code>med_acd</code>	<code>med_vcd</code>	<code>hs_mes_deporte</code>	<code>hs_mes_tv</code>	<code>padre_miopia</code>
	259	No	0.281	23.1	3.41	16.13	450	16	No

Para tratar este caso, se procedió a reemplazar el valor atípico por NA.

El dato observado se consideró implausible para el caso de estudio, por lo cual se decidió reemplazar el dato de esa variable por un NA. Para evitar errores en las estimaciones de los modelos se creó un subconjunto de datos (`datos_modelo`) que omite ese valor NA.

Se consideró hacerlo de esta manera para tener identificado explícitamente que valor fue considerado atípico.

2.4. Recodificación de Variables Categóricas

Se llevó a cabo una limpieza y recodificación de las variables categóricas (`miopia` y `padre_miopia`) para unificar respuestas:

- Se eliminaron espacios y se estandarizó el uso de mayúsculas y tildes ("sí", "si", "no", etc.).
- Se establecieron niveles fijos: "No" y "Sí", en ese orden, siendo "No" el nivel base para los modelos.

Estas tareas aseguran una correcta codificación y evitan errores en modelos estadísticos como regresión logística o Naive Bayes.

2.5. Clasificación de variables por tipo y rol

Luego del preprocesamiento, el dataset quedó conformado por 7 variables numéricas y 2 categóricas.

Las medidas optométricas se definieron de acuerdo a (Alrasheed & Aldakhil, 2022):

- AL (Axial Length)
- ACD (Anterior Chamber Depth)
- VCD (Vitreous Chamber Depth)

Por otro lado, el Spherical Equivalent (SE), variable `med_spheq`, se calcula como la esfera más la mitad del cilindro y representa una medida global del error refractivo (Enaholo, Musa, & Zeppieri, 2023).

A continuación, se clasifican según su tipo y rol en el análisis:

VARIABLE	TIPO	ROL	JUSTIFICACIÓN
id	Numérica	Excluida	Identificador sin valor predictivo.
miopia	Categórica (factor)	Respuesta	Es la variable objetivo: indica la presencia o ausencia de miopía.
med_spheq	Numérica	Predictora	Relacionada con el equivalente esférico de refracción ocular.
med_al	Numérica	Predictora	Longitud axial, fuertemente asociada a desarrollo de miopía.

med_acd	Numérica	Predictora	Profundidad de la cámara anterior del ojo.
med_vcd	Numérica	Predictora	Longitud vítrea, otro componente anatómico relevante.
hs_mes_deporte	Numérica	Predictora	Horas de actividad física; representa el estilo de vida.
hs_mes_tv	Numérica	Predictora	Horas de exposición a pantallas, posible factor de riesgo.
padre_miopia	Categórica (factor)	Predictora	Antecedente familiar, predictor importante según la literatura.

2.6. Resumen

El análisis exploratorio permitió:

- Comprender la estructura del dataset y limpiar las variables.
- Detectar y tratar un valor atípico extremo.
- Confirmar la ausencia de valores faltantes o duplicados relevantes (una vez tratados).
- Homogeneizar las categorías para su correcta interpretación.

Como resultado, se obtuvo un conjunto de datos limpio, consistente y listo para el análisis predictivo, respetando los requisitos de calidad para aplicar modelos como Regresión Logística y Naive Bayes.

3. Preparación del Dataset

3.1. Generación de la Partición Estratificada

Con el objetivo de ajustar y evaluar los modelos predictivos de manera adecuada, se procedió a dividir el conjunto de datos limpio (`datos_modelo`) en dos subconjuntos: entrenamiento (70%) y validación (30%).

Dado que la variable de respuesta miopía presenta un desbalance de clases (con aproximadamente un 13% de casos positivos y un 87% negativos), se implementó una partición estratificada utilizando la función `createDataPartition()`. Esta técnica garantiza que la proporción de clases se mantenga representada en ambos subconjuntos, evitando sesgos que puedan afectar el entrenamiento o la evaluación de los modelos.

Además, se redefinieron los niveles de las variables categóricas `miopia` y `padre_miopia`, estableciendo el nivel "No" como referencia. Esta codificación es relevante para la interpretación de los modelos estadísticos (por ejemplo, en regresión logística), aunque el interés principal del análisis está en detectar correctamente los casos positivos, es decir, los sujetos con miopía (nivel "Sí").

Resultado de la partición

- El conjunto de entrenamiento quedó conformado por 433 individuos, de los cuales 376 no presentan miopía y 57 sí la presentan.

- El conjunto de validación se compone de 184 individuos, con 160 casos sin miopía y 24 con miopía.

3.2 Distribución de clases por conjunto

La siguiente tabla resume la distribución de clases en ambos subconjuntos:

Distribución de la variable 'miopía' en los conjuntos de entrenamiento y validación

Conjunto	Clase	Cantidad	Porcentaje
Entrenamiento	No	376	86.8%
Entrenamiento	Sí	57	13.2%
Validación	No	160	87%
Validación	Sí	24	13%

3.3 Visualización

Para ilustrar gráficamente la distribución proporcional de la variable respuesta, se construyó un gráfico de barras apiladas. Este permite verificar que la proporción de clases se mantiene prácticamente constante entre los subconjuntos, garantizando que la validación sea representativa del problema real.

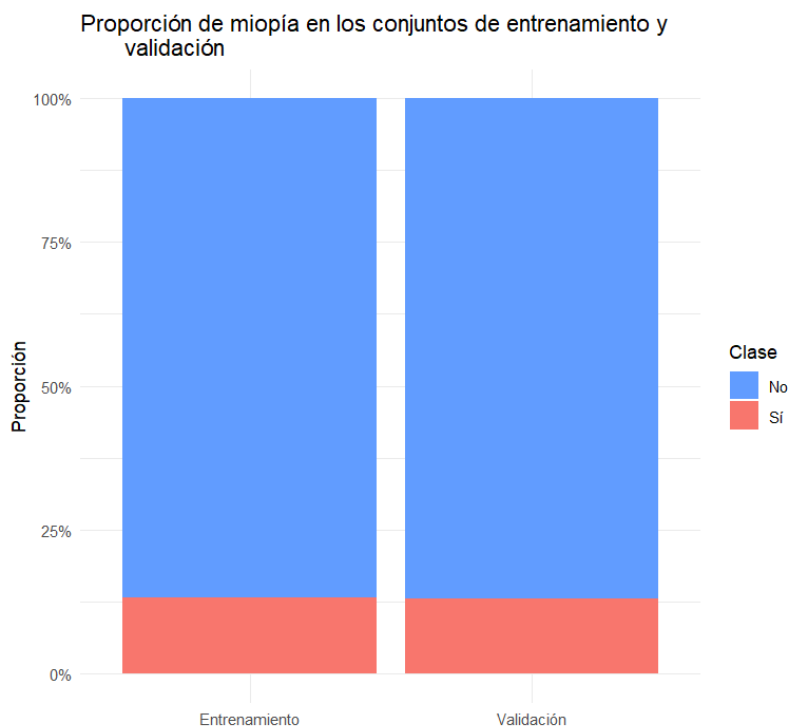


Ilustración 2 - Proporción de miopía en conjunto de entrenamiento y validación

4. Modelado con Regresión Logística

A continuación, se desarrollan tres modelos de regresión logística que relacionan la probabilidad de presentar miopía con diferentes combinaciones de variables disponibles en el conjunto de entrenamiento.

4.1. Modelo 1: Con Todas las Variables

A continuación, se desarrolla un modelo de regresión logística múltiple que incluye todas las variables independientes disponibles en el conjunto de entrenamiento: `med_spheq`, `med_al`, `med_acd`, `med_vcd`, `hs_mes_deporte`, `hs_mes_tv` y `padre_miopia`.

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= \log(odds) \\ &= \beta_0 + \beta_1 med_spheq + \beta_2 med_al + \beta_3 med_acd + \beta_4 med_vcd \\ &\quad + \beta_5 hs_mes_deporte + \beta_6 hs_mes_tv + \beta_7 padre_miopia\end{aligned}$$

Donde:

- p = es la probabilidad de tener miopía
- β_0 = intercepto del modelo
- β_i = coeficientes

Ecuación ajustada

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= 4.321 - 4.075 med_spheq + 0.009 med_al + 0.807 med_acd - 0.463 med_vcd \\ &\quad - 0.054 hs_mes_deporte + 0.009 hs_mes_tv + 0.724 padre_miopiaSi\end{aligned}$$

En la siguiente tabla se observa que, al nivel de significancia del 5%, las variables significativas son: `med_spheq`, `hs_mes_deporte` y `padre_miopia`. Siendo `med_spheq` la más significativa con un $p < 0,001$.

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR Z	PR(> z)	
intercepto	4.321214	8.718700	0.496	0.6202	> 0.05
med_spheq	-4.074980	0.549075	-7.422	1.16e-13	< 0.05
med_al	0.009131	1.305813	0.007	0.9944	> 0.05
med_acd	0.807326	1.326392	0.609	0.5427	> 0.05
med_vcd	-0.462875	1.222599	-0.379	0.7050	> 0.05
hs_mes_deporte	-0.054069	0.025290	-2.138	0.0325	< 0.05
hs_mes_tv	0.009439	0.033745	0.280	0.7797	> 0.05
padre_miopiaSi	0.723920	0.362455	1.997	0.0458	< 0.05

4.2. Modelo 2: Selección Automática (Stepwise)

Se aplicó la técnica de selección automática de variables mediante el criterio de información AIC, utilizando el método stepwise. Este procedimiento comienza con un modelo inicial y, en cada paso, evalúa la incorporación o eliminación de variables en función de su impacto en el valor del AIC.

El proceso identificó un modelo parsimonioso que incluye cuatro variables predictoras: `med_spheq`, `med_vcd`, `hs_mes_deporte` y `padre_miopiaSi`.

Ecuación ajustada

$$\log\left(\frac{p}{1-p}\right) = 6.439 - 4.120 \text{ med_spheq} - 0.390 \text{ med_vcd} - 0.052 \text{ hs_mes_deporte} + 0.709 \text{ padre_miopiaSi}$$

En la siguiente tabla se observa que, al nivel de significancia del 5%, las variables significativas son: `med_spheq`, `hs_mes_deporte` y `padre_miopiaSi`.

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR Z	Pr(> z)	
intercepto	6.43918	4.28528	1.503	0.1329	> 0.05
med_spheq	-4.12028	0.54838	-7.514	5.76e-14	< 0.05
med_vcd	-0.39053	0.27422	-1.424	0.1544	> 0.05
hs_mes_deporte	-0.05244	0.02494	-2.102	0.0355	< 0.05
padre_miopiaSi	0.70868	0.35883	1.975	0.0483	< 0.05

4.3. Modelo 3: Modelo Propuesto por el Grupo

Para la construcción del Modelo 3, se realizó una selección manual de variables combinando criterios estadísticos obtenidos en los modelos anteriores y el conocimiento teórico relacionado con los factores que pueden influir en la aparición de miopía en niños. Se seleccionaron las siguientes variables:

- `med_spheq`: variable clínica, altamente significativa en todos los modelos.
- `med_al`: longitud axial, fuertemente asociada a desarrollo de miopía, aunque en este caso no resultó estadísticamente significativa, se decidió incluirla por su sustento teórico.
- `med_vcd`: longitud vítrea, otro componente anatómico relevante.
- `padre_miopia`: variable familiar (herencia genética) que se mostró significativa.

Ecuación ajustada

$$\log\left(\frac{p}{1-p}\right) = 1.786 - 3.962 \text{ med_spheq} + 0.585 \text{ med_al} - 0.989 \text{ med_vcd} + 0.733 \text{ padre_miopiaSi}$$

En la siguiente tabla se observa que, al nivel de significancia del 5%, las variables significativas son: *med_spheq* y *padre_miopia*.

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR Z	Pr(> z)	
intercepto	1.7865	7.3794	0.242	0.8087	> 0.05
med_spheq	-3.9619	0.5266	-7.523	5.35e-14	< 0.05
med_al	0.5846	0.7960	0.734	0.4627	> 0.05
med_vcd	-0.9888	0.8175	-1.210	0.2264	> 0.05
padre_miopia	0.7331	0.3564	2.057	0.0397	< 0.05

4.4. Comparación y selección del modelo:

Se realizó un análisis de devianza para evaluar si la simplificación de los modelos implicaba una pérdida significativa en la capacidad explicativa. Como se puede observar, el *Modelo 2* no muestra una diferencia significativa en la devianza respecto al *modelo Completo* ($p = 0.787$), lo que indica que, a pesar de ser un modelo más sencillo, no pierde capacidad explicativa relevante.

MODELO	GRADOS DE LIBERTAD RESIDUAL	DEVIANZA RESIDUAL	CAMBIO DE GRADO DE LIBERTAD	CAMBIO DESVIACIÓN	Pr(>Chi)
M. 1 COMPLETO	425	219.69			
M. 2 SELECCIÓN AUTOMÁTICA	428	220.75	-3	-1.059	0.787
M. 3 PROPUESTO	428	225.00	0	-4.256	

El AIC se utilizó para evaluar el compromiso entre el ajuste del modelo y su complejidad. Se puede observar, el *Modelo 2* obtuvo el valor de AIC más bajo (230.75), lo que sugiere que es el modelo más eficiente y robusto.

MODELO	AIC
M. 1 COMPLETO	235.69
M. 2 SELECCIÓN AUTOMÁTICA	230.75
M. 3 PROPUESTO	235.00

Luego de comparar los tres modelos ajustados mediante el análisis de devianza y el AIC, se concluye que el *Modelo 2* (obtenido mediante selección automática por el método stepwise) es el más adecuado para explicar la presencia de miopía en la muestra analizada. Este modelo logra un equilibrio entre parsimonia y capacidad explicativa, eliminando variables que no aportan significancia al ajuste (como *med_al*, *med_acd* y *hs_mes_tv*) y conservando aquellas que sí resultaron ser estadísticamente significativas, como *med_spheq*, *hs_mes_deporte* y *padre_miopia*.

Además, el análisis de devianza mostró que este modelo no difiere significativamente en su ajuste respecto al modelo completo ($p = 0.787$), pero presenta un AIC más bajo (230.75 frente a 235.69 y 235.00), lo que lo convierte en una opción más eficiente y robusta.

Por estas razones, el *Modelo 2* se selecciona como el más adecuado para predecir la miopía infantil, permitiendo un modelo más simple sin perder capacidad predictiva.

5. Evaluación del Modelo Elegido

5.1. Significancia de Variables

Una vez seleccionado el *Modelo 2* mediante el criterio AIC y el análisis de devianza, se procedió a evaluar la significancia estadística de cada una de las variables incluidas en el modelo, a través de pruebas de hipótesis individuales para los coeficientes.

La hipótesis nula y alternativa para cada variable son:

- **H₀:** El coeficiente de la variable es igual a cero (la variable no tiene efecto significativo sobre la probabilidad de presentar miopía).
- **H₁:** El coeficiente de la variable es distinto de cero (la variable tiene efecto significativo sobre la probabilidad de presentar miopía).

El nivel de significancia adoptado es $\alpha = 0.05$. A continuación se resumen los resultados obtenidos para cada variable del modelo:

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR Z	Pr(> z)	
intercepto	6.43918	4.28528	1.503	0.1329	> 0.05
med_spheq	-4.12028	0.54838	-7.514	5.76e-14	< 0.05
med_vcd	-0.39053	0.27422	-1.424	0.1544	> 0.05
hs_mes_deporte	-0.05244	0.02494	-2.102	0.0355	< 0.05
padre_miopiaSi	0.70868	0.35883	1.975	0.0483	< 0.05

Interpretación de resultados

- **med_spheq** resultó altamente significativa ($p < 0.001$), con un coeficiente negativo ($\beta = -4.12$). Esto indica que a mayor valor de **med_spheq** (menor miopía o más hipermetropía), disminuye fuertemente la probabilidad de que el niño tenga miopía. Este análisis se alinea con la literatura clínica (Chen, Liu, Sha, Yang, & Yu, 2021).
- **hs_mes_deporte** también fue significativa ($p = 0.0355$), con un coeficiente negativo: ($\beta = -0.052$). Esto sugiere que una mayor cantidad de horas dedicadas al deporte disminuye la probabilidad de miopía, lo que podría vincularse a hábitos más saludables y menor tiempo frente a pantallas.
- **padre_miopiaSi** también resultó significativa ($p = 0.0483$), con un coeficiente β positivo, lo cual confirma la influencia hereditaria en la aparición de miopía infantil.
- **med_vcd** no fue significativa ($p = 0.1544$), aunque se mantuvo en el modelo por su posible valor explicativo dentro del ajuste global.

En síntesis, según el test de Wald, tres de las cuatro variables incluidas en el modelo presentan un efecto estadísticamente significativo sobre la probabilidad de desarrollar miopía.

5.2. Test de Bondad de Ajuste

Para evaluar la adecuación del modelo de regresión logística ajustado (*Modelo 2*) a los datos, se aplicó la prueba de bondad de ajuste de Hosmer-Lemeshow. Esta prueba compara las frecuencias observadas y esperadas de la variable respuesta agrupando los datos en deciles según las probabilidades predichas por el modelo.

Planteo de hipótesis

- H_0 : el modelo se ajusta bien a los datos; no existen diferencias significativas entre las probabilidades observadas y las predichas.
- H_1 : El modelo no se ajusta bien a los datos; existen diferencias significativas entre las probabilidades observadas y las predichas.

Los resultados obtenidos son los siguientes:

- Estadístico Chi-cuadrado: 6.9316
- Grados de libertad: 8
- p-valor: 0.544

Dado que el p-valor = 0.544 es mayor que el nivel de significación ($\alpha = 0.05$), no se rechaza la hipótesis nula. Esto indica que no hay evidencia estadística para afirmar que el modelo se ajusta mal a los datos.

Por lo tanto, se concluye que el *Modelo 2* presenta un ajuste adecuado según este test.

5.3. Interpretación de Coeficientes

Se seleccionó el coeficiente de la variable `padre_miopia` del *Modelo 2* para su interpretación en términos de odds.

- Coeficiente estimado (`padre_miopiaSí`): 0.709

Interpretación en términos de odds

La estimación del coeficiente para la variable `padre_miopiaSí` fue $\beta = 0.709$, con un p-valor de 0.048, lo que indica una relación estadísticamente significativa al 5%.

En términos de odds, esto implica que tener al menos un progenitor con miopía se asocia con un incremento en las probabilidades relativas (odds) de que el niño también tenga miopía, comparado con un niño sin antecedentes familiares, manteniendo constantes las demás variables del modelo.

El odds ratio (OR) correspondiente es:

$$OR = e^{0.709} \approx 2.03$$

Esto significa que los niños que tienen un padre miope tienen el doble de chances (2.03 veces más) de presentar miopía en relación a quienes no tienen dicho antecedente, en iguales condiciones clínicas y ambientales.

Este hallazgo refuerza la influencia del componente hereditario en el desarrollo de la miopía infantil.

5.4. Análisis de Multicolinealidad

La multicolinealidad ocurre cuando dos o más variables explicativas están fuertemente relacionadas entre sí. Esto dificulta saber cuál de ellas realmente está afectando a la variable que queremos predecir, y puede hacer que los coeficientes del modelo sean inestables o poco confiables.

Para verificar la presencia de multicolinealidad entre las variables explicativas, se analiza el Factor de Inflación de la Varianza (VIF) para cada variable independiente.

Interpretación del VIF

De acuerdo con los valores obtenidos, como se puede ver en la siguiente tabla:

TÉRMINO	VIF
med_spheq	1.064715
med_vcd	1.029287
hs_mes_deporte	1.041326
padre_miopia	1.003535

- Todos los VIF son cercanos a 1 y muy por debajo del umbral comúnmente utilizado de 5 o 10, que indicaría un posible problema de multicolinealidad.
- Estos resultados sugieren que no existen problemas de multicolinealidad entre las variables regresoras del modelo.

5.5. Influencia de Casos

Sobre el modelo seleccionado por selección automática de variables (med_spheq, med_vcd, hs_mes_deporte y padre_miopia), se evaluó la presencia de observaciones influyentes utilizando la Distancia de Cook.

Distancia de Cook

El gráfico de Distancia de Cook permitió identificar algunos casos que superan levemente el umbral teórico de influencia ($4/n$). Aunque podrían considerarse potencialmente influyentes, ninguno

presenta un valor lo suficientemente elevado como para sugerir un impacto desmedido en los parámetros del modelo.

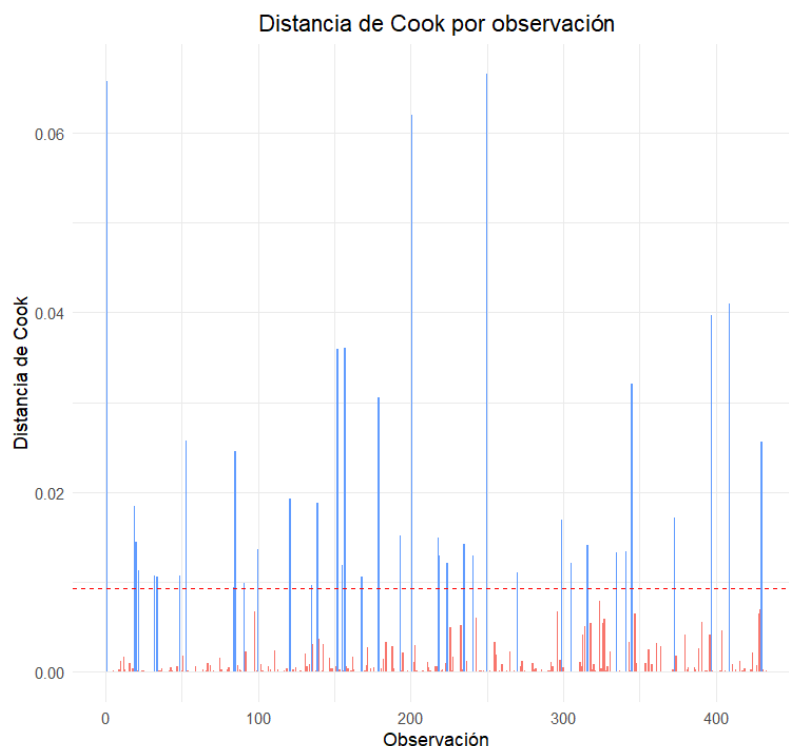


Ilustración 3 - Distancia de Cook por observación

Decisión sobre mantener/eliminar observaciones

Tras analizar el gráfico de distancia de Cook, se concluye que no existen observaciones con una influencia sustancial sobre los resultados del modelo. Por ello, se decidió mantener todas las observaciones en el análisis, dado que la eliminación de casos con influencia mínima no aportaría una mejora significativa en la estabilidad ni en la capacidad predictiva del modelo.

Asimismo, no se identificaron errores de carga o datos atípicos relevantes. Mantener todas las observaciones permite conservar la variabilidad natural de la muestra y evita el riesgo de sobreajuste (overfitting) o de obtener modelos excesivamente depurados, que podrían limitar la generalización del modelo a nuevos datos.

5.6. Síntesis del Modelo Seleccionado

El *Modelo 2*, obtenido mediante selección stepwise, fue elegido como el más adecuado por presentar:

- Una estructura simple y parsimoniosa, con menos variables pero buen desempeño.
- Un ajuste satisfactorio a los datos, respaldado por un AIC reducido y buena discriminación.
- Predictores estadísticamente significativos, especialmente `med_spheq` y `padre_miopia`.

Además, superó favorablemente los tests de ajuste y diagnóstico, incluyendo el test de Hosmer-Lemeshow, la evaluación de multicolinealidad y el análisis de observaciones influyentes.

6. Modelado con Naive Bayes

6.1. Ajuste del modelo

Para complementar el análisis predictivo, se ajustó un modelo de clasificación Naive Bayes, utilizando las variables seleccionadas previamente en el *Modelo 2* de regresión logística:

- `med_speq`
- `med_vcd`
- `hs_mes_deporte`
- `padre_miopia.`

Estas variables fueron seleccionadas por su relevancia clínica y estadística, y permiten mantener una comparación coherente entre ambos enfoques de modelado.

Antes del ajuste se confirmó que:

- La variable `miopia` está codificada como factor binario, con niveles "No" y "Sí", siendo "Sí" la clase positiva de interés.
- Las variables numéricas cumplen con el supuesto de normalidad condicional que asume el clasificador Naive Bayes para modelar la distribución dentro de cada clase.
- La variable `padre_miopia`, de tipo categórico, es tratada de forma automática mediante estimación de probabilidades condicionales simples.

El modelo fue ajustado sobre el conjunto de entrenamiento (`train`) utilizando la función `naiveBayes()`. Las probabilidades a priori fueron:

- $P(\text{miopiaNo}) = 0.868$
- $P(\text{miopiaSí}) = 0.132$

El resumen del modelo mostró las siguientes probabilidades condicionales y medias por clase:

Distribución de variables numéricas según la clase
(Naive Bayes)

Variable	Clase	Media	Desvío Estándar
med_spheq	No	0.910	0.612
med_spheq	Sí	0.197	0.418
med_vcd	No	15.388	0.677
med_vcd	Sí	15.384	0.611
hs_mes_deporte	No	11.896	7.259
hs_mes_deporte	Sí	10.316	8.227

Probabilidades condicionales de miopía según
padre miope (Naive Bayes)

Padre miope	P(miopía = No)	P(miopía = Sí)
No	54.3%	31.6%
Sí	45.7%	68.4%

6.2. Validación del modelo en el conjunto test

Una vez ajustado, el modelo se evaluó sobre el conjunto de validación (*test*). Se realizaron las siguientes acciones:

- Se obtuvo la clasificación predicha para cada individuo del conjunto *test*.
- Se construyó la matriz de confusión comparando las predicciones con los valores reales.
- A partir de esta matriz se calcularon las métricas estándar de desempeño: *accuracy*, *sensibilidad* y *especificidad*.

		Predicción	
		No	Sí
Real	No	158	2
	Sí	21	3

A partir de esta matriz se calcularon las siguientes métricas:

Métricas de evaluación del
modelo Naive Bayes sobre el
conjunto de test

Métrica	Valor
Exactitud (Accuracy)	87.5%
Sensibilidad (Recall)	12.5%
Especificidad	98.8%
Valor Predictivo Positivo	60.0%
Valor Predictivo Negativo	88.3%
Índice Kappa	0.170
Exactitud Balanceada	55.6%

Estas métricas indican que el modelo tiene un excelente desempeño para identificar a los niños sin miopía (alta especificidad), pero baja capacidad para detectar los casos con miopía (sensibilidad baja). Esto es habitual en contextos con clases desbalanceadas, como el presente caso.

6.3. Curva ROC y AUC

Además de la clasificación binaria, se estimaron las probabilidades de pertenencia a la clase "Sí", necesarias para calcular la curva ROC (Curva Característica Operativa del Receptor) y el AUC (Área Bajo la Curva), que permiten evaluar el desempeño global del modelo sin depender de un punto de corte específico.

- La curva ROC mostró un comportamiento aceptable en términos de discriminación.
- El valor de **AUC = 0.852**, indica una buena capacidad de separación global entre clases.

La curva ROC que se presenta a continuación, donde en el eje X se representa la especificidad (proporción de verdaderos negativos), y en el eje Y, la sensibilidad (proporción de verdaderos positivos).

En este caso, se observa que la curva se mantiene elevada respecto de la diagonal, lo que indica un buen poder discriminativo del modelo. Esto se corresponde con un AUC superior a 0.80, lo que significa que el modelo es eficaz para distinguir entre individuos con y sin miopía.

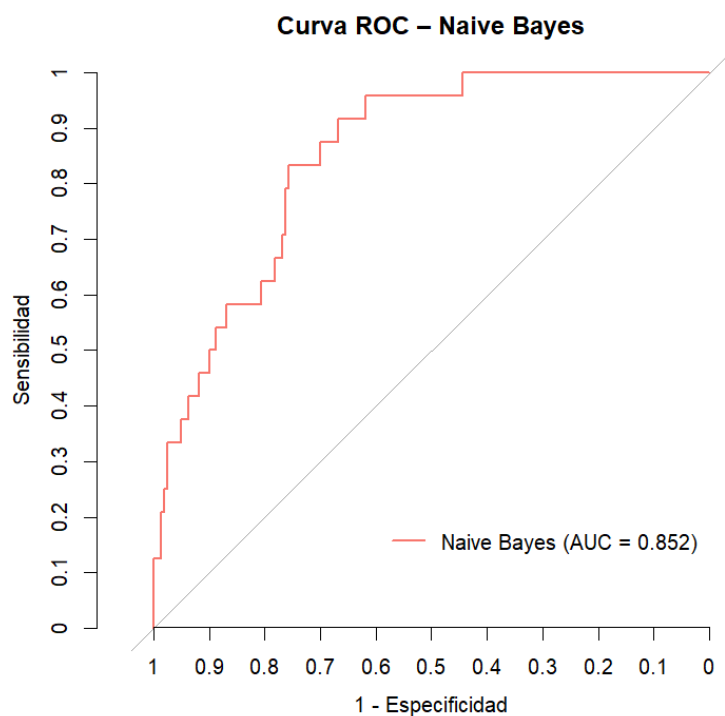


Ilustración 4 - Curva ROC - Naive Bayes

6.4. Síntesis interpretativa

El *Modelo Naive Bayes*, ajustado con las mismas variables del *Modelo 2* de regresión logística, demostró un desempeño mixto sobre el conjunto *test*:

- Su exactitud global (87.5%) es buena, pero esto se debe principalmente a su alta especificidad (98.8%).
- La capacidad para detectar niños con miopía fue muy baja (sensibilidad del 12.5%), lo que implica un riesgo alto de falsos negativos si el objetivo fuera capturar todos los casos positivos.
- No obstante, su AUC elevado (0.852) indica que, en términos generales, el modelo tiene una buena capacidad discriminativa, por lo que un ajuste del umbral de decisión podría mejorar la sensibilidad si fuera necesario.

7. Evaluación y Comparación de Modelos en el Conjunto Test

A continuación, se presentan los resultados de evaluación del *Modelo 2* y el *Modelo Naves Bayes* aplicados al conjunto *test*, incluyendo la curva ROC, el AUC y las métricas de desempeño derivadas de la matriz de confusión.

7.1. Curva ROC y AUC

Se calcularon las probabilidades predichas para el conjunto *test*, y se construyó la curva ROC correspondiente. El Área Bajo la Curva (AUC) obtenida fue:

- *Modelo 2* = 0.873
- *Modelo Naves Bayes* = 0.852

En ambos casos el AUC indica una muy buena capacidad discriminativa del modelo para distinguir entre niños con y sin miopía, ya que un valor cercano a 1 refleja una buena separación entre clases; en estos casos, los modelos logran un equilibrio razonable entre sensibilidad y especificidad. Siendo el *Modelo 2* el de mejor desempeño.

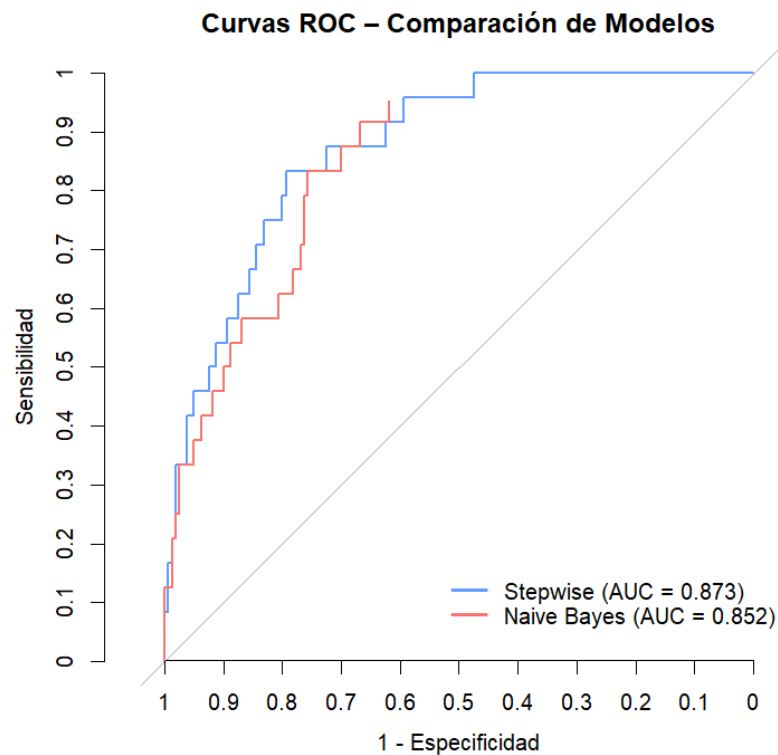


Ilustración 5 - Curvas ROC - comparación de Modelos

7.2. Tabla de Clasificación y Métricas

Métricas

Las métricas de cada uno de los modelos son las siguientes:

Comparación de métricas entre Modelo Stepwise y Naive Bayes

Modelo	Exactitud	Kappa	Sensibilidad	Especificidad	VPP	VPN	AUC
Logística Stepwise	89.1%	0.442	41.7%	96.2%	62.5%	91.7%	87.3%
Naive Bayes	87.5%	0.170	12.5%	98.8%	60.0%	88.3%	85.2%

Matriz de confusión

La matriz de confusión para cada uno de los modelos fue la siguiente:

Modelo 2 - Stepwise			
		Predicción	
		No	Sí
Real	No	154	6
	Sí	14	10

Modelo Naive Bayes			
		Predicción	
		No	Sí
Real	No	158	2
	Sí	21	3

7.4. Discusión de Resultados

De acuerdo a la tabla comparativa de métricas sobre el conjunto *test*, donde se puede observar que el *Modelo 2* presentó mejor rendimiento global, destacándose especialmente en sensibilidad (41.7% vs. 12.5%) y Kappa (0.442 vs. 0.170), esto indica una mejor capacidad para detectar casos positivos de miopía y una mayor concordancia con la clasificación real, respectivamente.

Al comparar ambas matrices de confusión, se observa que el *Modelo 2* logra detectar una mayor cantidad de niños con miopía (10 casos frente a 3 del *Modelo Naive Bayes*), lo que refleja una mayor sensibilidad (41.7% vs. 12.5%). Sin embargo este mejor desempeño se da a costa de un leve aumento en los falsos positivos (6 en el *Modelo 2* vs 2 en el *Modelo Naive Bayes*). En contraste, el *Modelo Naive Bayes* muestra mayor especificidad (98.8%), pero falla en identificar correctamente a la mayoría de los niños con miopía.

8. Elección del Modelo Según el Objetivo

8.1. Elección del Punto de Corte Óptimo

El modelo seleccionado, *Modelo 2* de regresión logística con método Stepwise, calcula la probabilidad de que un niño presente miopía. Para convertir estas probabilidades en una decisión binaria ("Sí" o "No"), es necesario definir un punto de corte o umbral.

Si bien el umbral tradicional de referencia es 0.5, este puede ser subóptimo en contextos donde las clases están desbalanceadas o cuando se prioriza un tipo específico de error (falsos positivos o falso negativos).

Por ello, se evaluó el desempeño del *Modelo 2* utilizando una secuencia de umbrales que varían en incrementos de 0.04, con el objetivo de explorar cómo cambian las métricas de desempeño (*sensibilidad, especificidad, precisión*) según el punto de corte.

Se destacan tres umbrales representativos:

- **Umbral 0.1:** es un punto de corte bajo, que maximiza la sensibilidad del modelo, permitiendo detectar la mayor cantidad de casos de niños con miopía, a costa de una mayor tasa de falsos positivos.
- **Umbral 0.5:** busca un equilibrio entre sensibilidad y especificidad, pero puede ser inadecuado en contextos con clases desbalanceadas o si se prioriza un objetivo específico.
- **Umbral 0.62:** este punto de corte ofrece una mayor exactitud global sobre el conjunto de validación, priorizando la correcta clasificación total aunque con menor sensibilidad.

La siguiente tabla muestra la evolución de las métricas del *Modelo 2* a medida que se modifica el punto de corte. Se observa que, como es esperable, a menor umbral aumenta la sensibilidad pero disminuye la especificidad y la precisión.

Comparación de métricas para distintos umbrales de clasificación

Umbral	Accuracy	Precision	Sensibilidad	Especificidad	F1
0.10	0.723	0.304	0.875	0.700	0.452
0.14	0.777	0.351	0.833	0.769	0.494
0.18	0.804	0.375	0.750	0.812	0.500
0.22	0.821	0.400	0.750	0.831	0.522
0.26	0.832	0.405	0.625	0.863	0.492
0.30	0.837	0.412	0.583	0.875	0.483
0.34	0.853	0.448	0.542	0.900	0.491
0.38	0.864	0.480	0.500	0.919	0.490
0.42	0.870	0.500	0.458	0.931	0.478
0.46	0.886	0.579	0.458	0.950	0.512
0.50	0.891	0.625	0.417	0.963	0.500
0.54	0.886	0.600	0.375	0.963	0.462
0.58	0.891	0.667	0.333	0.975	0.444
0.62	0.897	0.727	0.333	0.981	0.457
0.66	0.897	0.727	0.333	0.981	0.457
0.70	0.886	0.667	0.250	0.981	0.364
0.74	0.880	0.667	0.167	0.988	0.267
0.78	0.886	0.800	0.167	0.994	0.276
0.82	0.880	1.000	0.083	1.000	0.154
0.86	0.875	1.000	0.042	1.000	0.080
0.90	0.870	NA	0.000	1.000	NA

8.2. Comparación del Modelo 2 con distintos umbrales y Naive Bayes

A continuación, se compararon las métricas de desempeño del *Modelo 2* bajo los tres puntos de corte mencionados y las del *Modelo Naive Bayes*, con el objetivo de evaluar cuál se ajusta mejor a cada necesidad específica:

Comparación de métricas: Modelo Logístico (en distintos umbrales) vs Naive Bayes

Modelo	Umbral	Exactitud	Kappa	Sensibilidad	Especificidad	VPP	VPN	F1	AUC
Logística Stepwise (0.10)	0.1	72.3%	0.320	87.5%	70.0%	30.4%	97.4%	45.2%	87.3%
Logística Stepwise (0.50)	0.5	89.1%	0.442	41.7%	96.2%	62.5%	91.7%	50.0%	87.3%
Logística Stepwise (0.62)	0.62	89.7%	0.409	33.3%	98.1%	72.7%	90.8%	45.7%	87.3%
Naive Bayes	-	87.5%	0.170	12.5%	98.8%	60.0%	88.3%	20.7%	85.2%

El siguiente gráfico muestra cómo varían la sensibilidad, la especificidad y la precisión a medida que se modifica el punto de corte en el *Modelo 2*. Podemos observar que, a medida que el umbral es más bajo, la sensibilidad es mayor, mientras que la especificidad y la precisión disminuyen, y viceversa.

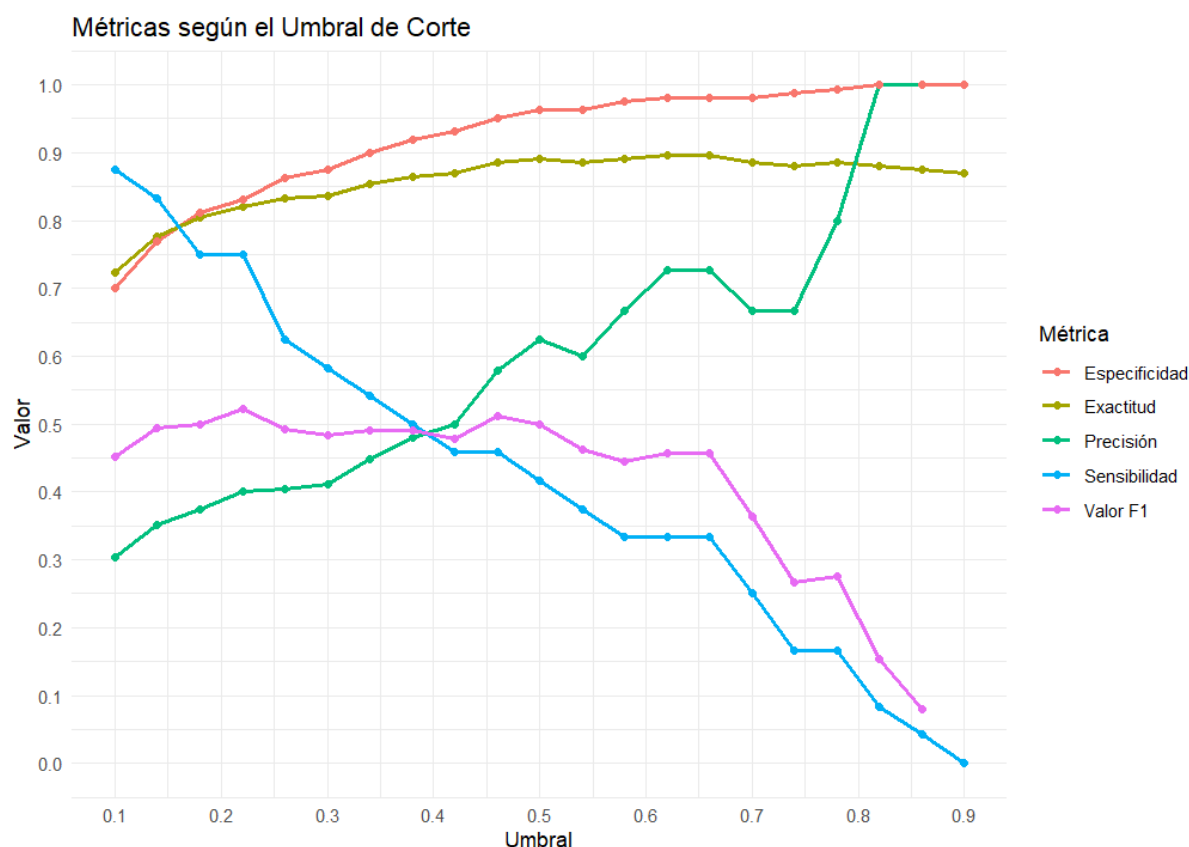


Ilustración 6: Métricas según el umbral de corte

Los principales hallazgos son los siguientes:

- El *Modelo 2* con umbral 0.1 se destaca por su alta sensibilidad (87.5%), lo que lo convierte en la mejor opción cuando el objetivo es detectar la mayor cantidad de niños con miopía, aunque esto implique menor precisión y más falsos positivos.
- El *Modelo 2* con umbral 0.62 alcanza la mayor exactitud global (89.7%) y especificidad (98.1%), siendo ideal cuando se busca minimizar los errores globales y clasificar correctamente la mayoría de los casos, aunque con una sensibilidad baja (33.3%) para detectar miopía.
- El *Modelo 2* con umbral 0.5 se mantiene como opción intermedia sin destacarse en ninguna métrica particular.
- El *Modelo Naive Bayes* mostró un AUC aceptable (85.2%) y buena especificidad (98.8%), pero una sensibilidad muy baja (12.5%), lo que lo hace menos adecuado para detectar casos positivos en este contexto.

Comparación de Desempeño de Modelos de Clasificación en Métricas Normalizadas

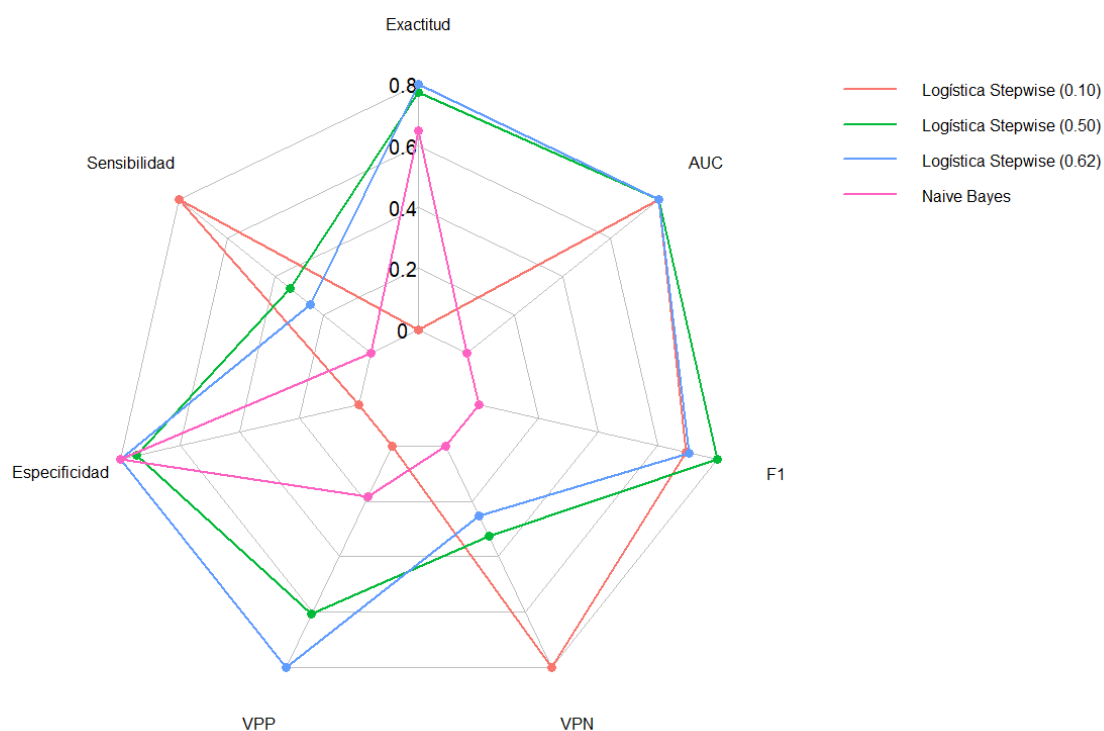


Ilustración 7: Comparación de Desempeño de Modelos de Clasificación mediante Métricas Normalizadas

8.3. Elección del Modelo para Detectar la Mayor Cantidad de Niños con Miopía

Cuando el objetivo es priorizar la detección de la mayor cantidad de niños con miopía, se busca un modelo que tenga alta sensibilidad, es decir capacidad de identificar correctamente los casos positivos, aun cuando esto implique aceptar un mayor número de falsos positivos.

En este caso, el *Modelo 2* con un umbral de 0.1 resulta ser la opción más adecuada. Este modelo logra una sensibilidad del 87.5%, esto significa que permite detectar casi 9 de cada 10 niños que efectivamente presentan miopía. Si bien la exactitud global disminuye y la especificidad es más baja en este escenario, esto es aceptable ya que en este contexto sería más grave no identificar a un niño con miopía (error tipo II) que clasificar incorrectamente a un niño sin la afección (error tipo I).

Por lo tanto, para un objetivo preventivo y clínico, donde es esencial no dejar casos sin detectar, se recomienda utilizar el *Modelo 2* con punto de corte 0.1.

8.4. Elección Modelo para Minimizar Errores Globales

Cuando el objetivo principal es minimizar la cantidad total de errores de clasificación, es decir, obtener la mayor exactitud global posible y reducir tanto los falsos positivos como los falsos negativos, se debe priorizar la precisión general del modelo.

En este caso, el *Modelo 2* con un umbral de 0.62 demostró tener el mejor desempeño en términos de exactitud (89.7%) y especificidad (98.1%), junto con un índice Kappa de 0.409, lo que refleja una buena concordancia general. Este punto de corte permite un balance adecuado entre sensibilidad y especificidad, pero está especialmente enfocado en minimizar los errores totales, siendo menos permisivo en la detección de casos positivos marginales.

Por lo tanto, si el objetivo es obtener un modelo robusto y preciso en términos globales, sin enfocarse exclusivamente en la detección de positivos, se recomienda el uso del *Modelo 2* con umbral de 0.62.

9. Conclusiones Finales

Este trabajo permitió construir y comparar modelos predictivos para detectar miopía infantil a partir de variables clínicas y de estilo de vida.

El *Modelo 2* de regresión logística seleccionado mostró un mejor desempeño y mejor capacidad para identificar casos positivos frente al *Modelo Naive Bayes*. A su vez se analizó su desempeño considerando diferentes puntos de corte (umbrales de 0.1, 0.5 y 0.62) para observar su comportamiento.

Se concluye que:

- El *Modelo 2*, con cualquiera de los umbrales analizados, tiene mejor desempeño que el *Modelo Naive Bayes*, tanto para detectar la mayor cantidad de casos de miopía como para minimizar los errores globales.
- Si el objetivo es detectar la mayor cantidad de casos de miopía, se recomienda el *Modelo 2* con umbral 0.1.
- Si se busca minimizar los errores globales, la mejor opción es el *Modelo 2* con umbral 0.62.

Como posibles mejoras al presente trabajo, en un futuro se podría incorporar validación cruzada, para una selección de modelos más robusta y obtener una estimación más confiable del error de predicción. También sería recomendable aplicar técnicas de balanceo de clases durante el entrenamiento (como SMOTE) para mejorar la detección de casos positivos en conjuntos desbalanceados.

Además, se podría explorar la inclusión de términos de interacción, como la posible combinación entre horas de deporte y padres con miopía, lo que podría revelar efectos combinados relevantes.

Finalmente, se podría complementar el análisis con diagnósticos de residuales más detallados, como gráficos de residuales de Pearson o devianza frente a las probabilidades predichas, para evaluar mejor los supuestos del modelo de regresión logística.

10. Anexos

Se adjuntan los archivos correspondientes al Trabajo Práctico para facilitar su revisión y replicación:

- **Código fuente en R:** archivo `TP_Final_grupo_10.R`, que contiene todo el script utilizado para el análisis, limpieza de datos, modelado y visualización.
- **Código fuente en formato HTML:** archivo `TP_Final_grupo10_script.html`, generado a partir del script R Markdown, que presenta el desarrollo completo, los gráficos y los resultados del trabajo.
- **Código fuente en formato PDF:** archivo `TP_Final_grupo10_script.pdf`, también generado a partir del script R Markdown, que presenta el desarrollo completo, los gráficos y los resultados del trabajo.

Estos materiales permiten reproducir el análisis y verificar los resultados presentados en este informe.

11. Referencias

- Alrasheed, S. H., & Aldakhil, S. (21 de 11 de 2022). *The Open Ophthalmology Journal*. Obtenido de Corneal Curvature, Anterior Chamber Depth, Lens Thickness, And Vitreous Chamber Depth: Their Intercorrelations With Refractive Error In Saudi Adults:
https://openophthalmologyjournal.com/VOLUME/16/ELOCATOR/e187436412209300/FULLTEXT/?utm_source=chatgpt.com
- Chen, S., Liu, X., Sha, X., Yang, X., & Yu, X. (13 de 11 de 2021). *National Library of Medicine*. Obtenido de Relationship between axial length and spherical equivalent refraction in Chinese children:
https://pmc.ncbi.nlm.nih.gov/articles/PMC10577836/?utm_source=chatgpt.com
- Enaholo, E. S., Musa, M. J., & Zeppieri, M. (11 de 05 de 2023). *National Library of Medicine*. Obtenido de The Spherical Equivalent:
https://www.ncbi.nlm.nih.gov/books/NBK589657/?utm_source=chatgpt.com