CIMAT

Análisis de Artículos Científicos

Introducción a Ciencia de Datos

Autores:
María Alejandra Borrego Leal
Iván García Mestiza
Rodolfo de Jesús Ramírez Lucario

Profesor:
Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas, A. C.

13 de Octubre de 2025

1. Introducción

En el presente reporte analizaremos dos artículos que usan herramientas de regresión, tanto lineal como logística, describiendo los modelos utilizados, interpretando los resultados obtenidos y realizando alternativas para el tratamiento de los datos.

El primer artículo analizado es [Williams et al., 2021], el cual tiene como objetivo determinar si el peso corporal del planeador de Leadbeater (*Gymnobelideus leadbeateri*) se relaciona con variables ambientales y con el sexo del individuo. Dicho artículo realiza el ajuste de un modelo de regresión lineal, previamente seleccionando dos variables de un total de 29 y descartando el resto. De esta manera concluye que el sexo y la altitud de la región geográfica en donde habita esta especie permiten explicar el peso de los individuos. Este artículo fue escogido puesto que la selección de variables fue de manera "empírica", escogiendo varias de ellas y aplicando un test de correlación de Pearson, y aunque implementaron selección *stepwise*, no mostraron de manera clara los resultados obtenidos que sustentaran trabajar con únicamente dos variables. En este proyecto implementaremos *elastic net* y regresión con datos agrupados dependiendo del lugar en donde se hizo la muestra, y con esto obtendremos modelos que explican de una mejor manera los datos.

El segundo artículo analizado es [Rimal et al., 2025], el cual realiza un análisis comparativo de distintos modelos, incluyendo regresión logística, para predecir enfermedades cardiacas, incluyendo factores como edad, sexo, antecedentes médicos, entre otros. Dicho artículo fue seleccionado puesto que, aunque el análisis de la regresin logística fue bueno, mostrando distintas métricas de evaluación, no realizó una exploración más profunda como para seleccionar variables, lo cual es de suma importancia, puesto que esto ayudaría a saber cuáles factores son los que tienen más peso en el desarrollo de enfermedades cardiacas. Por dicha razón, en este proyecto implementaremos dos propuestas de selección de variables: usando modelos bayesianos y *stepwise*, lo cual mejora la interpretabilidad de los resultados puesto que se reduce el número de variables involucradas, sin sacrificar el poder predictivo del modelo.

2. Análisis del artículo "Relationship between body weight and elevation in Leadbeater's possum (Gymnobelideus leadbeateri)"

El presente artículo indica que el tamaño corporal en los mamíferos está determinado por una combinación de factores evolutivos, fisiológicos, ecológicos y morfológicos. Analizar estas variaciones permite comprender mejor los procesos que influyen en la morfología animal. En este estudio se evaluó si el peso corporal del planeador de Leadbeater (*Gymnobelideus leadbeateri*) se relaciona con variables ambientales y con el sexo del individuo. Para ello, se aplicaron modelos de regresión lineal considerando variables geográficas (como latitud y altitud), factores locales asociados a la productividad del bosque (tipo de vegetación, pendiente, orientación y humedad topográfica), y la variable sexo.

2.1. Descripción de las variables

El conjunto de datos para el análisis incluyó 135 mediciones de peso corporal del planeador de Leadbeater (52 hembras y 83 machos), recolectadas de 63 cajas nido distribuidas a lo largo del rango de localización de la especie. Para las covariables se incluyeron mediciones topográficas y el tipo de bosque como variables explicativas potenciales en el modelado de la variación del peso corporal, considerando un total de 29 variables independientes para el modelo. La descripción de cada una de las varibales en el conjunto de datos lo podemos ver en la Tabla 1.

Variable	Descripción
pointid	Identificador del individuo
Date	Fecha de captura del espécimen
Month	Mes de captura del espécimen
Colony	Identificador de la colonia: La especie vive en colonias matriarcales de hasta 12 individuos, dominadas por una hembra adulta reproductora
NestBox	Identificador de caja de captura
Retrap / new	Indicador de si el individuo ya había sido capturado antes o no
Sex	Sexo del individuo
Wgt	Peso del individuo en gramos
Age	Edad del individuo
Breeding status	Indicador de si el espécimen se reproduce
Pouch	Estado de la cría
Location	Lugar de captura
Forest_type	Tipo de bosque
Old_growth	Bosque maduro o primario
east	Coordenadas proyectadas en la dirección Este
north	Coordenadas proyectadas en la dirección Norte
zone	Zona
hemisphere	Hemisferio
lat	Latitud del lugar de captura
lon	Longitud del lugar de captura
psf	Percent Slope Factor: Mide el porcentaje de inclinación del terreno
grid_conv	Convergencia de los meridianos en la proyección cartográfica (diferencia angular entre el norte del mapa y el norte geográfico)
DEM	Altitud del terreno
Slope	Pendiente, inclinación del terreno
Aspect	Dirección cardinal hacia la que se inclina la pendiente del terreno
TWI	Índice topográfico de humedad; combina pendiente y área de drenaje acumulada.
Relative Slope Index 10m Resolution	Mide cuán empinado es el sitio relativo a su entorno inmediato
Valley Depth 10 m Resolution	Cuánto está “hundido” el punto respecto a su entorno
Slope Length Factor 10 m Resolution	Describe la tendencia del agua a acumularse mientras viaja hacia abajo de una pendiente a lo largo de cierta longitud
Leo location	Localización del individuo (nombre del lugar)

Tabla 1: Descripción de variables.

2.2. Análisis y replicación de resultados

El artículo comienza su análisis estadístico realizando una serie de pruebas de Pearson entre algunas covariables y la variación en el peso corporal del animal. Este último se define como

$$\text{Variación del Peso Corporal} = \text{Peso Máximo - Peso Mínimo.}$$

La primera prueba que se realiza es entre el número de individuos capturados en cada caja y la variación del peso corporal. La segunda es entre el número de individuos capturados en cierto mes y la variación del peso. Finalmente, se realiza el test entre el número de individuos recolectados en cada tipo de bosque y la variación del peso corporal. Los resultados de las prueba se muestran en la Tabla 2.

Variable	r	t	df	P
Netboxes vs. weight variation	0.741	8.54	60	0.0000
Month vs. weight variation	0.349	1.05	8	0.3228
Forest type vs. weight variation	0.047	0.07	2	0.9525

Tabla 2: Test de Pearson entre el número de animales capturados por caja contra la variación del peso, el número de animales capturados por mes contra la variación del peso y el número de animales capturados por tipo de bosque contra la variación del peso.

La primera variable no es una variable biológica explicativa real (no representa un factor ecológico), sino un artefacto del esfuerzo de muestreo y además tiene una alta correlación con la variable independiente. Consideramos que por estas razones, los autores decidieron eliminarla de las variables predictoras, lo mismo ocurre con la variable del mes. En el caso de la variable de tipo de Bosque, esta sí representa una variable biológica explicativa real además de que por sí sola no tiene una alta capacidad de predicción, suponemos que por esta razón los autores decidieron incluirla en la selección de variables.

Implícitamente se asume que todas las hipótesis del modelo de regresión lineal se satisfacen, pero en el artículo se hacen ciertas simplificaciones para la disminución del número de variables, y posteriormente se lleva a cabo el ajuste del modelo. Se realizó una selección de variables mediante el método Stepwise eligiendo como métrica de error el AIC. Las variables que compitieron fueron *lat, dem, slope, aspect, twi, sex* y el modelo que resultó ganador fue aquel con las variables *dem* y *sex*. Las métricas de este modelo se encuentran en la Tabla 3. Podemos ver que las covariables seleccionadas son significativas para la regresión además de que la prueba F salió significativa i.e. se rechazó la hipótesis nula de que todos los coeficientes sean 0. Por otro lado, el p-valor de la prueba de Jarque-Bera nos indica que no se rechaza que los datos sean Normales.

Variable	**Coef.**	**Std. Error**	t	$P > \|t\|$	**[0.025**	**0.975]**
Intercept	108.776	4.476	24.301	0.000	99.922	117.630
Sex (male)	3.680	1.625	2.264	0.025	0.465	6.894
Elevation (*dem*)	0.0195	0.003	5.791	0.000	0.013	0.026
Model summary:						
R-squared = 0.216	Adj. R-squared = 0.204	F(2,132) = 18.21	Prob(F) = 1.03×10^{-7}			
AIC = 983.6	BIC = 992.3	Log-Likelihood = -488.80				
Durbin–Watson = 1.79	Jarque–Bera $p = 0.0565$	$n = 135$				

Tabla 3: Ordinary Least Squares (OLS) para el peso (wgt) como una función del sexo (sex) y elevación (dem).

Se comenta que el modelo se evaluó con gráficos de errores pero no se presentan. Para la validación de este modelo, realizamos el análisis en R y se obtuvieron los gráficos de residuos de la Figura 1.

De la Figura 1 podemos concluir lo siguiente:

- Cómo puede verse en el gráfico **Residuals vs Fitted**, la relación entre las covariables y la variable de respuesta pereciera ser lineal dado que los errores se encuentran al rededor del 0.

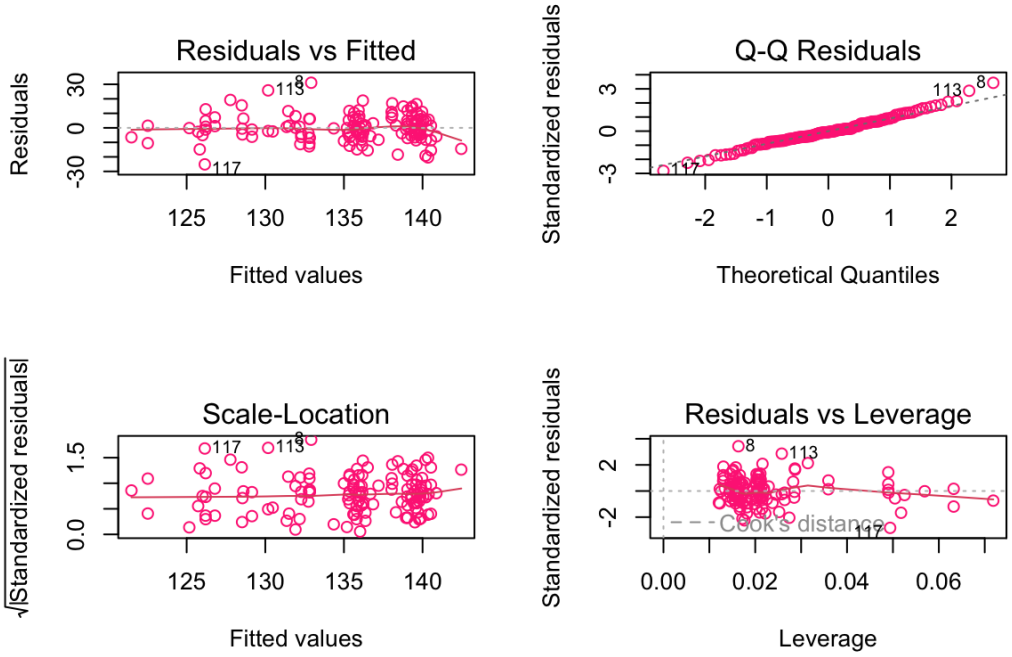

Figura 1: Gráficas de validación del modelo de regresión lineal ajustado.

- En el gráfico **Q-Q Residuals** notamos que los puntos se encuentran al rededor de la recta identidad lo que sugiere que los errores estandarizados tienen una distribución normal.
- La gráfica **Scale-Location** sugiere que los errores son homocedásticos pues la varianza del error permanece constante.
- En la Gráfica de la distancia Cook, que podemos ver en la Figura 2 indica que no hay observaciones influyentes en la regresión pues todos los valores estan por debajo del 0.5.

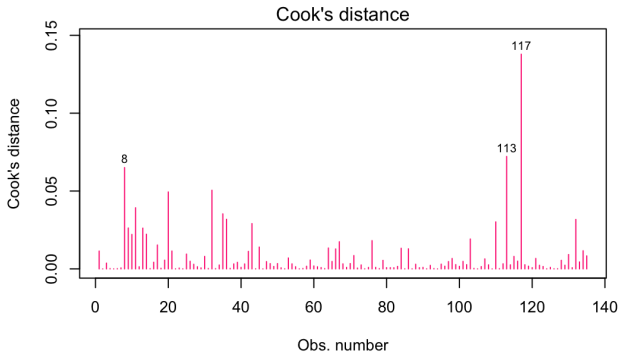

Figura 2: Distancia de Cook en el modelo de regresión lineal ajustado.

Los coeficientes de la regresión nos indican que existe una correlación positiva entre el sexo y el peso i.e. que si es macho, el peso medio del individuo sube 3.68 gramos. De igual manera, entre más alto

fue encontrado el individuo, mayor peso este tiene esto último, como comentan los autores del artículo, es consistente con la regla de Bergmann la cuál indica que: el tamaño del cuerpo del individuo está negativamente correlacionada con la temperatura. Así, dado que entre más alto se está sobre el nivel de mar menor es la temperatura, hemos encontrado que la Regla de Bergmann se satisface.

Podemos encontrar las rectas de regresión separadas por sexo junto con los intervalos de confianza de la predicción media en la Figura 3.

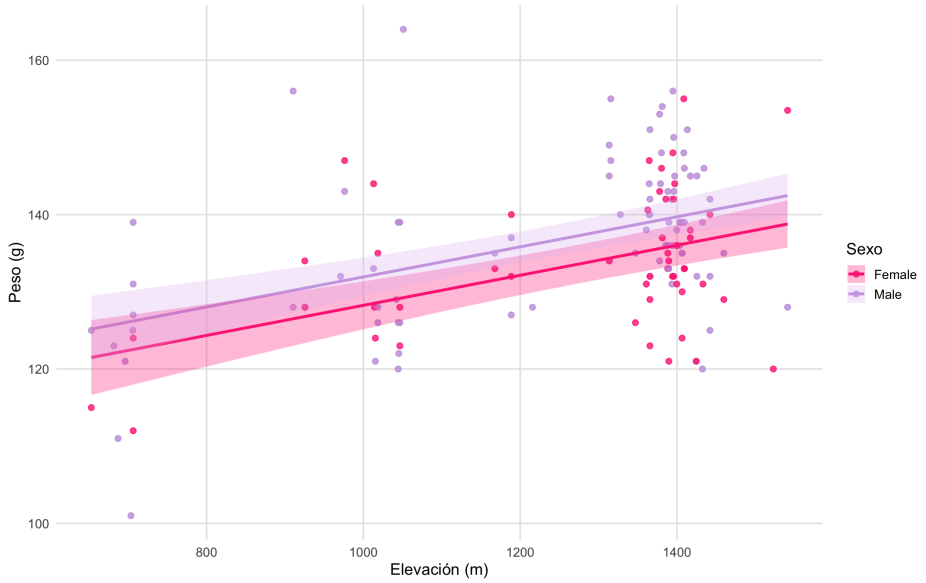

Figura 3: Modelo de regresión lineal ajustado por sexo con bandas de predicción de la media.

Los creadores del artículo utilizaron un módulo de *R* llamado *emmeans* que sirve para calcular las Medias Marginales estimadas. Las Medias Marginales Estimadas son el valor promedio que predice el modelo para un grupo específico (por ejemplo, machos o hembras) controlando las demás covariables por la media. Primero se calcularon las Medias Marginales Estimadas del sexo, los resultados se observan en la tabla 4. Esta tabla nos indica que el peso marginal promedio de los machos es 137 con un intervalo de confianza del 95 % de 135 y 139, en el caso de las hembras, el peso marginal promedio es de 133 con un intervalo de confianza del 95 % de 131 y 136.

Sex	EMMean	SE	df	Lower CL	Upper CL
Female	133	1.27	132	131	136
Male	137	1.01	132	135	139

Tabla 4: Medias Marginales Estimadas (EMMs) del peso por sexo.

2.3. Exploración de alternativas

En el ajuste de la regresión lineal propuesta por el artículo se observa que, a pesar de tener un valor pequeño de la estadística R^2, los supuestos de la regresión parecen cumplirse, lo cual se puede comprobar en el análisis de la Figura 1. Esta situación es común en ámbitos de biología, en donde las mediciones normalmente poseen mucho ruido, por lo que la medición de la R^2 y R^2 ajustada no son consideradas como definitivas para la evaluación de los modelos. Sin embargo, como se comentó previamente, la selección de variables careció de generalidad en cierta manera, y no se consideraron técnicas como la regularización para selección de variables.

Por lo anterior, como primera propuesta realizamos regresión con regularización, la cual es útil cuando se tienen variables predictoras altamente correlacionadas, hace la inferencia más robusta y menos sensible a la multicolinealidad, y los coeficientes ajustados tienden a ser más consistentes. Por lo general, las regresiones Ridge y Lasso se utilizan de manera separada para la reducción de variables. Por un lado, la regresión Ridge es útil para situaciones de multicolinealidad, y disminuye los valores de los coeficientes de manera suave, aunque no permite una selección de variables como tal. Por otro lado, la regresión Lasso da la oportunidad de que algunos coeficientes se vuelvan cero, permitiendo una selección de variables automática, aunque es inestable si las variables predictoras están altamente correlacionadas. De esta manera surge la regresión *Elastic Net* [Zou and Hastie, 2005], la cual combina las penalizaciones L_1 y L_2 de los métodos Lasso y Ridge, respectivamente, y balancea la selección de variables con la estabilidad. Formalmente, el estimador de *Elastic Net* es el que minimiza la función

$$\mathrm{EN}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - (X\boldsymbol{\beta})_i\right)^2 + \alpha\lambda \sum_{j=1}^{k} |\beta_j| + \alpha(1-\lambda) \sum_{j=1}^{k} |\beta_j|^2 .$$

A evaluar la Regresión *Elastic Net* sobre una rejilla de pesos para α y λ obtenemos que se obtiene el mejor ajuste para $\alpha = 1$ y $\lambda = 1$, lo que equivale al caso de la Regresión Lasso. Es decir, en este caso los problemas de multicolinealidad no afectan a la selección de variables que realiza Lasso, así que esta termina siendo la mejor opción. Para dicha regresión, la Tabla 5 muestra las variables seleccionadas así como los coeficientes ajustados. Además, para dicho modelo se obtuvo una R^2 ajustada de 0.1786, la cual sigue siendo buena en comparación con el ajuste realizado en el artículo.

Variable	**Coeficiente ajustado**
Elevation (*dem*)	3.226
Latitude (*lat*)	0.368
Percent slope factor (*psf*)	0.281

Tabla 5: Coeficientes ajustados en el mejor modelo de Regresión *Elastic Net*.

Como podemos ver en la Tabla 5, la elevación sigue siendo un factor importante en el peso, pero este modelo no considera que el sexo sea lo suficientemente significativo, y en cambio involucra la ubicación geográfica y la pendiente del terreno. Por lo tanto, este modelo indica que las variables geográficas son las que más influyen para el peso, lo cual tiene sentido, puesto que los animales se deben adaptar a su entorno de la manera en que más les convenga. En la Figura 4 se muestra la gráfica de residuales del modelo ajustado. Como podemos ver, aunque logramos reducir el número de variables con éxito, los errores no parecen ser homocedásticos, por lo que en este caso se pierde un poco la validez del modelo ajustado. Más aún, el AIC obtenido fue de 990.353, y el BIC resultó igual a 1001.9746, ambos mayores que los del modelo original, por lo que surge la necesidad de explorar métodos alternativos.

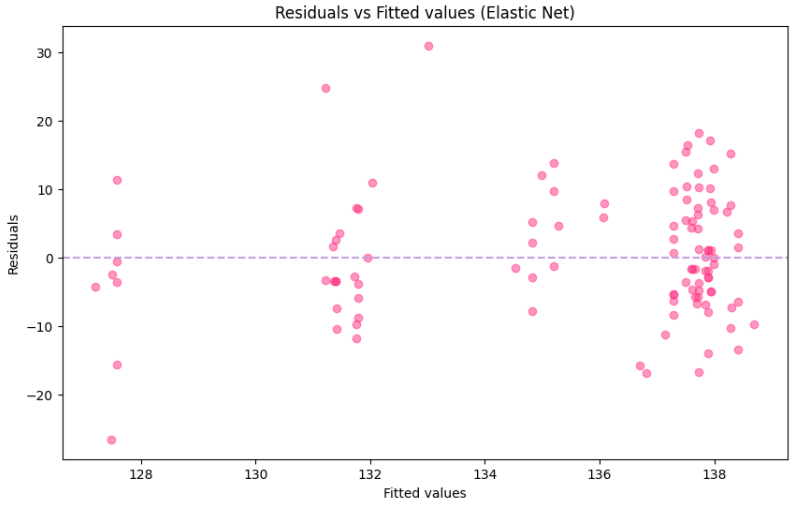

Figura 4: Residuales del modelo *Elastic Net* ajustado.

Para el análisis del planeador de Leadbeater (*Gymnobelideus leadbeateri*), los individuos fueron capturados y colocados en diferentes cajas-nido (*nest box*), cada una con condiciones controladas. El artículo menciona que bajo una exposición prolongada de los animales a estas cajas-nido, su temperatura resulta ser un factor influyente para el tamaño de los individuos, puesto que la termo-regulación puede requerir distintos niveles de energía dependiendo de la temperatura, y esto a su vez afecta al tamaño y peso de los animales. Por consiguiente, un análisis alternativo que se puede dar a los datos consiste en regresión con datos agrupados de acuerdo a estas cajas-nido. Puesto que animales de la misma zona geográfica terminan en la misma caja-nido si se encuentran cercanos entre sí, tiene sentido que esta variable de cierta forma resuma al resto de variables geográficas. Los resultados obtenidos por medio de dicha regresión se encuentran en la Tabla 6.

Parameter	**Coef.**	**Std.Err.**	**z**	**P$>\|z\|$**	**[0.025**	**0.975]**
Intercept	106.564	5.402	19.726	0.000	95.976	117.152
sex[T.Male]	3.711	1.404	2.644	0.008	0.960	6.462
Elevation (*dem*)	0.021	0.004	5.094	0.000	0.013	0.030
Group Var	32.870	2.453				

Model Information	
No. Observations:	135
No. Groups:	62
Min. group size:	1
Max. group size:	7
Mean group size:	2.2
Method:	REML
Scale:	55.9824
Log-Likelihood:	-488.2549
Converged:	Yes

Tabla 6: Resultados de la Regresión agrupada por medio de la variable *nestbox*.

Como se puede observar en la Tabla 6, hay 62 cajas-nido, y tanto el sexo como la elevación vuelven a resaltar como variables influyentes. Además, en este modelo se obtuvo una R^2 igual a 0.599, que es significativamente mayor a los obtenidos previamente, lo cual ayuda a reafirmar la validez de esta metodología. Por otro lado, el AIC y BIC, respectivamente, fueron iguales a 986.509 y 1001.036. Esto también representa un incremento con respecto a los originales. Sin embargo, al analizar la gráfica de los residuales, que se muestra en la Figura 5, se puede notar que los errores se distribuyen de manera homogénea a lo largo de todo el gráfico, e incluso se comportan de mejor manera que el modelo propuesto por el artículo original, lo cual reafirma la validez de este método.

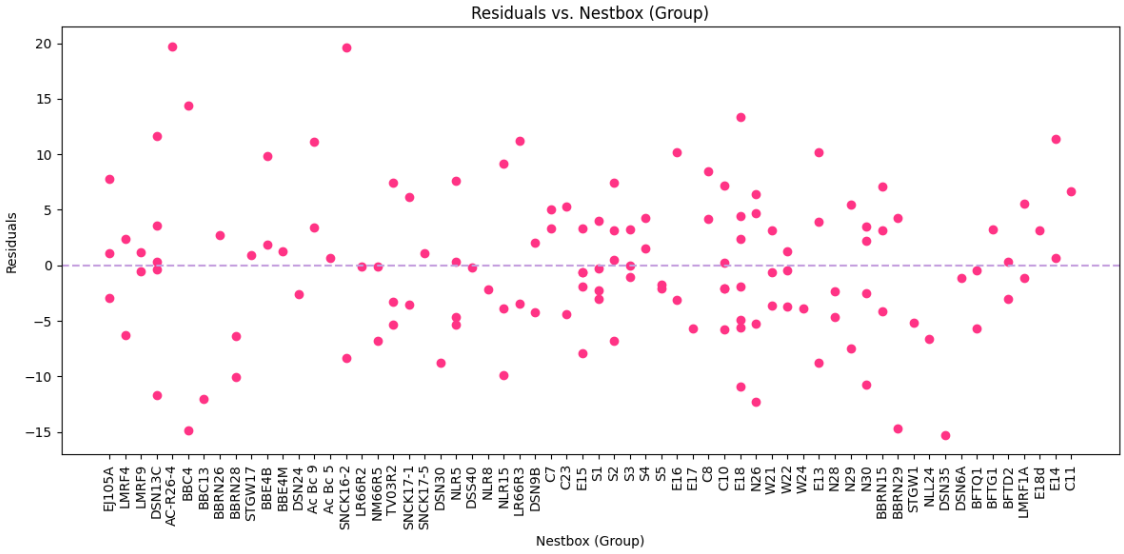

Figura 5: Residuales del modelo de regresión agrupada ajustado.

Por lo tanto, podemos ver que existe una buena concordancia entre el análisis realizado en el artículo original con los métodos propuestos, lo cual refuerza la robustez de los resultados obtenidos y nos permite concluir que, en efecto, tanto el sexo como la ubicación geográfica son influyentes para determinar el peso de los individuos. Además, aunque las metodologías aquí planteadas son un poco más generales y robustas, se obtienen conclusiones similares, por lo que, si bien la selección de variables realizada por el artículo original pudo no ser del todo satisfactoria en su justificación, logró el objetivo planteado.

3. Análisis del artículo "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross - validation for improved accuracy"

En este artículo se realiza un análisis comparativo de distintos modelos utilizando cuatro metodologías de validación cruzada: Regresión Logística (LR), Máquina de Vectores de Soporte (SVM), K-Vecinos Más Cercanos (KNN) y Bosque Aleatorio (RF), aplicadas a conjuntos de datos abiertos sobre enfermedades cardiacas. El objetivo es identificar fácilmente la precisión promedio de las predicciones de enfermedades del corazón de los distintos modelos y, posteriormente, hacer recomendaciones para la selección de estos.

Para el análisis realizado de este artículo nos enfocamos solamente en la regresión logística que presenta dicho estudio. Nos limitamos a este modelo con el propósito de explorar a detalle la manera en que se plantea, sus supuestos y su desempeño en la predicción de enfermedades cardiacas.

3.1. Descripción de Variables

El estudio adopta un enfoque riguroso de preparación de datos y validación de modelos. El conjunto de datos, compuesto por 13 variables predictoras y una variable objetivo binaria, pasa por un preprocesamiento, donde se utiliza la codificación one-hot encoding para trabajar con las variables categóricas y métodos de reescalamiento para las numéricas. Se trabaja entonces con un total de 8 variables categóricas y 5 numéricas, que se presentan descritas en la Tabla 7.

Variable	Descripción
age	Edad del paciente en años
sex	Sexo del paciente
cp	Tipo de dolor de pecho reportado
trestbps	Presión arterial en reposo (mm Hg)
chol	Colesterol sérico en la sangre (mg/dL)
fbs	Indicadora de si el nivel de glucosa en la sangre es mayor a cierto umbral (binaria)
restecg	Resultados de electrocardiograma en reposo (categórica)
thalach	Frecuencia cardiaca máxima alcanzada
exang	Indicadora si tiene una angina inducida por el ejercicio (binaria)
oldpeak	Depresión ST inducida por ejercicio
slope	Pendiente del segmento ST máximo
ca	Número de vasos principales coloreados mediante fluoroscopia
thal	Estatus de talasemia (categórica)

Tabla 7: Descripción de variables.

Posteriormente, se utiliza una partición entrenamiento-prueba del 80:20 con estratificación para mantener la proporción de clases, seguida de un procedimiento de validación cruzada que se emplea para graficar curvas de aprendizaje, garantizando una evaluación integral del rendimiento del modelo.

3.2. Análisis y replicación de resultados

Una vez realizado el preprocesamiento de los datos, el análisis comienza obteniendo la correlación entre la variable dependiente y las variables independientes utilizando un mapa de calor, esto para entender qué variables están significativamente relacionadas entre sí y la fuerza de esta relación respecto a la

probabilidad de causar enfermedad cardiaca. Al replicar el análisis del artículo, generamos el siguiente mapa de calor, el cual coincide con el presentado en el estudio original.

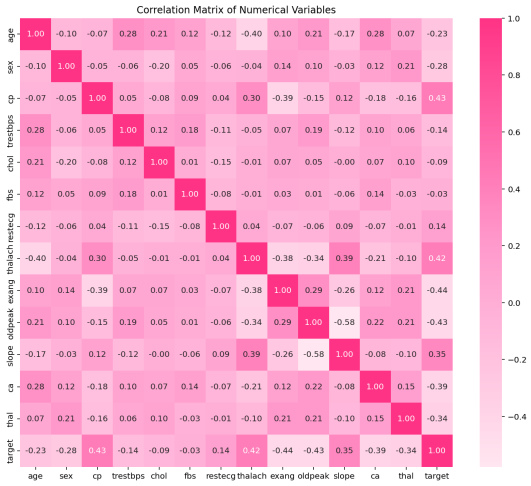

Figura 6: Mapa de calor de correlación entre las variables numéricas.

Posteriormente, se utiliza el modelo de regresión logística configurado con un máximo de mil iteraciones para entrenar sobre los datos. Para ello, se divide el conjunto de datos en subconjuntos de entrenamiento y prueba y se ajusta el modelo a los datos de entrenamiento. Una vez ajustado el modelo, se aplicó validación cruzada de 5 iteraciones, cambiando los conjuntos de prueba en cada iteración. Se realizaron predicciones sobre cada conjunto de prueba y se calcularon métricas de evaluación tales como: accuracy, precision, recall y F1-score. Presentamos en la Tabla 8 los resultados obtenidos una vez replicado el análisis descrito.

Métrica	**Iteración 1**	**Iteración 2**	**Iteración 3**	**Iteración 4**	**Iteración 5**
Accuracy	0.8852	0.8525	0.7705	0.8167	0.8667
Precision	0.8824	0.8529	0.7317	0.7895	0.8378
Recall	0.9091	0.8788	0.9091	0.9091	0.9394
F1	0.8955	0.8657	0.8108	0.8451	0.8857

Tabla 8: Métricas de evaluación del modelo.

En el artículo original se muestran los valores de las métricas de evaluación correspondientes a las iteraciones con mayor y menor nivel de accuracy que se obtuvieron. Comparándolas con las que obtuvimos en nuestra réplica, donde la iteración 1 corresponde al mayor accuracy y la iteración número 3 al menor, notamos que obtuvimos resultados muy cercanos: en la mayor iteración obtuvimos un 88.5 % de accuracy contra un 88 % del mostrado en el artículo y en la menor iteración obtuvimos un 77 % de accuracy contra un 78 % del estudio. Al igual que el artículo, consideramos la iteración con mayor nivel de accuracy del modelo en las iteraciones realizadas para obtener la correspondiente curva ROC y matriz de confusión.

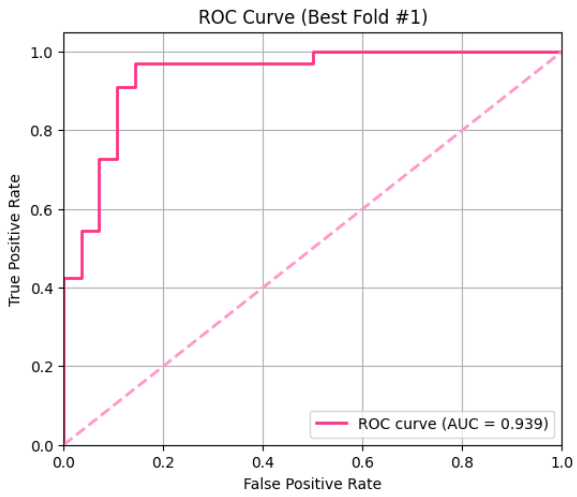

Figura 7: Curva ROC.

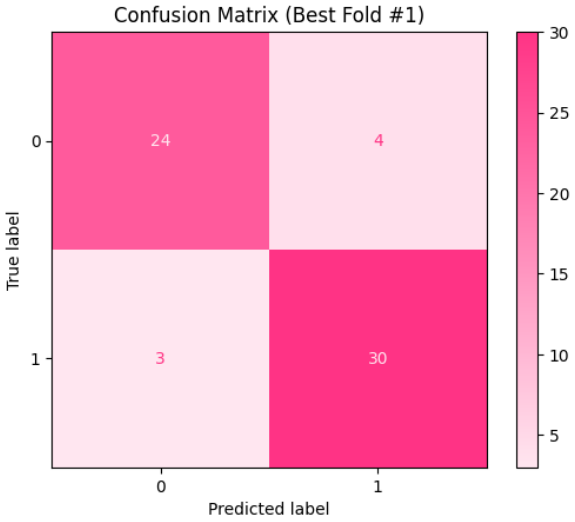

Figura 8: Matriz de Confusión.

En la Figura 7 podemos observar la curva ROC en donde apreciamos que la mejor exactitud que alcanzó el modelo arroja un valor AUC de 0.939, siendo prácticamente el mismo que se obtuvo en el estudio original con un valor AUC de 0.94. En la Figura 8 vemos la matriz de confusión en la que apreciamos que el modelo predice correctamente cerca del 89 % de los casos, indicando así un desempeño de clasificación satisfactorio.

Las curvas de aprendizaje suelen generarse mediante métodos como la búsqueda en malla con validación cruzada, que evalúa al modelo a lo largo de un rango de parámetros, aprovechando cantidades absolutas de grupos de entrenamiento para graficar y analizar las tendencias de rendimiento.

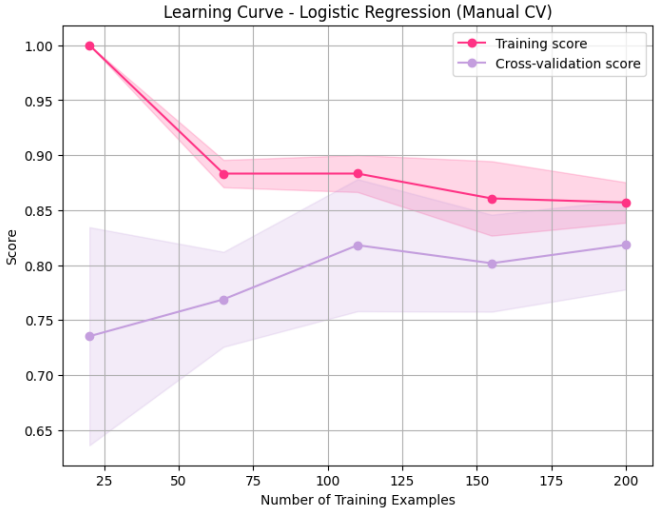

Figura 9: Curvas de aprendizaje del modelo.

En la Figura 9 las dos líneas representan la curva de validación, que varía de forma gradual, siendo la línea inferior la que indica el error de entrenamiento o la puntuación de precisión. Esta curva muestra cómo cambian las métricas de error conforme se incrementan los conjuntos de entrenamiento y validación,

hasta que el modelo alcanza su mejor ajuste. Cada línea describe los efectos combinados del modelo con los conjuntos de datos cardiacos. Observamos cómo el puntaje de entrenamiento comienza cercano a 1, disminuyendo de manera gradual conforme aumenta el número de muestras de entrenamiento, lo que indica que el modelo generaliza mejor a medida que dispone de más datos. Por otro lado, la curva de validación cruzada muestra una mejora progresiva en la capacidad predictiva del modelo.

Comparando la Figura 9 con la curva de aprendizaje reportada en el artículo apreciamos comportamientos muy similares con una disminución del puntaje de entrenamiento y un aumento sostenido de la validación cruzada conforme crece el tamaño de los datos de entrenamiento. La principal diferencia radica en que, en nuestro resultado, las curvas presentan una ligera pero mayor dispersión y una brecha algo más pronunciada, lo que podría deberse a variaciones en el muestreo o en la configuración del número de iteraciones y particiones de la validación cruzada. Sin embargo, en general la tendencia coincide fuertemente con la reportada en el artículo, lo que confirma la validez de nuestra réplica.

En términos generales, la réplica del análisis de regresión logística que llevamos a cabo reproduce de manera consistente los resultados reportados en el artículo original. Las métricas de desempeño presentan variaciones mínimas, dentro del margen esperado al considerar diferencias en la partición aleatoria de los datos o en los procedimientos internos de validación cruzada. La ligera diferencia en los valores máximos y mínimos de accuracy no implica un cambio importante en la capacidad predictiva del modelo, lo que demuestra que nuestra réplica logra reproducir con éxito los resultados del artículo.

3.3. Exploración de alternativas

Si bien logramos replicar de buena manera los resultados del artículo, sin mostrar mucha discrepancia y confirmando la validez de sus conclusiones con respecto al análisis de regresión logísitica, en este estudio no buscan encontrar el modelo más óptimo (o simple), pues solamente están comparando el modelo completo (considerando las 13 variables predictoras) con otros modelos. Por esta razón, proponemos utilizar un par de métodos para realizar una buena selección de modelo.

Como primera propuesta, realizamos un análisis con modelos bayesianos que nos permite cuantificar la incertidumbre de los parámetros de manera explícita, proporcionando distribuciones posteriores para cada coeficiente en lugar de estimaciones puntuales. Asumimos que la variable objetivo sigue una distribución Bernoulli cuyo parámetro es la probabilidad de que haya presencia de enfermedad cardiaca. Asignamos tanto al intercepto como los coeficientes distribuciones normales centradas en cero con desviación estándar igual a 10, esto para no imponer una restricción fuerte a priori. El análisis de las distribuciones posteriores de los coeficientes nos permite identificar las variables cuya masa posterior se concentra lejos de cero, lo cual sugiere un efecto relevante en la predicción del evento. Para visualizar estos resultados presentamos las Figuras 10 y 11.

La Figura 10 muestra los intervalos de credibilidad del 94 % para cada parámetro, y en la Figura 11 podemos apreciar las densidades posteriores superpuestas, que reflejan la magnitud y dirección de cada efecto. Entonces, basándonos en los resultados obtenidos (y sabiendo que cada coeficiente representa las variables predictoras en el orden en que se encuentran en el dataframe), las variables que no se consideran significativas para nuestra selección del modelo son: *age*, *fbs*, *restecg*, *slope*, *trestbps* y *chol*. De manera que el mejor modelo tiene como variables predictores a *exang*, *ca*, *oldpeak*, *cp*, *thal*, *sex* y *thalach*, dónde hemos descartado aquellas en las que el 0 cayó en el intervalo de credibilidad de los coeficientes.

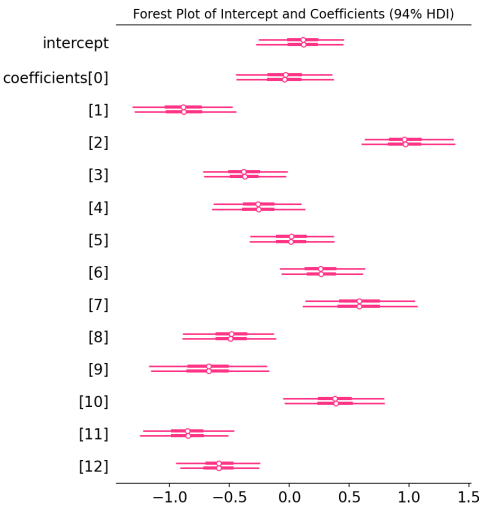

Figura 10: Curva ROC.

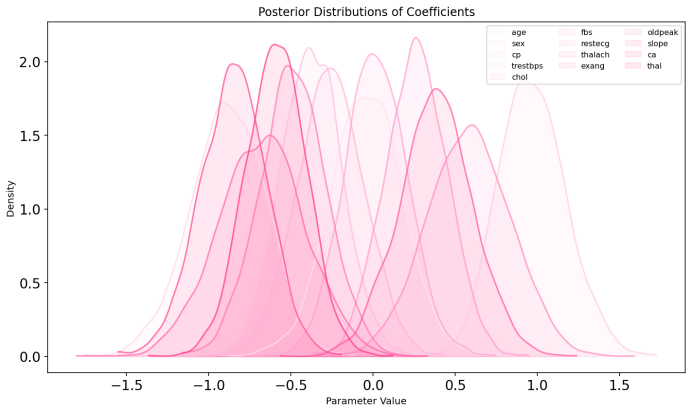

Figura 11: Matriz de Confusión.

Para complementar nuestra propuesta utilizamos el método de selección de variables por pasos (stepwise). El proceso comienza creando 13 modelos de regresión, cada uno de los cuales utiliza exactamente una de las variables. Posteriormente, la importancia de las variables se clasifica según su capacidad individual para explicar la variación en el resultado. Presentamos en la Tabla 9 el resultado obtenido al aplicar este método.

Model Summary		
Dep. Variable:	target	No. Observations: 303
Model:	Logit	Df Residuals: 295
Method:	MLE	Df Model: 7
Pseudo R-squ.:	0.4653	Log-Likelihood: -111.66
LL-Null:	-208.82	LLR p-value: 1.824e-38
Converged:	True	Covariance Type: nonrobust

Variable	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.4636	1.482	0.313	0.754	-2.440	3.367
exang	-1.0447	0.389	-2.686	0.007	-1.807	-0.282
ca	-0.7133	0.174	-4.090	0.000	-1.055	-0.372
oldpeak	-0.7406	0.182	-4.061	0.000	-1.098	-0.383
cp	0.7872	0.175	4.505	0.000	0.445	1.130
thal	-0.8963	0.275	-3.265	0.001	-1.434	-0.358
sex	-1.3896	0.406	-3.425	0.001	-2.185	-0.594
thalach	0.0237	0.009	2.685	0.007	0.006	0.041

Tabla 9: Resultados de la Regresión logística con el método *stepwise*.

Notamos que el modelo seleccionado con este método considera las mismas variables predictoras que las seleccionadas con el análisis bayesiano. Para validar la selección de este modelo presentamos en la Figura 12 la gráfica del cuadrado de los residuales vs el leverage (o influencia). En ella observamos como la mayoría de las observaciones están concentradas cercas al eje vertical, con bajos valores tanto de leverage como de residuales, lo cual indica que el modelo no presenta problemas importantes de ajuste

ni observaciones altamente influyentes. Podemos apreciar solamente ciertos puntos alejados o con valores de leverage que superan la línea morada que marca el promedio, pero ninguno con nivel que sugiera un impacto desproporcionado en la estimación de los coeficientes. Esto respalda la estabilidad del modelo ajustado y sugiere que la selección de variables mediante el método stepwise condujo a una especificación adecuada, sin evidencia de observaciones que comprometan la validez del ajuste.

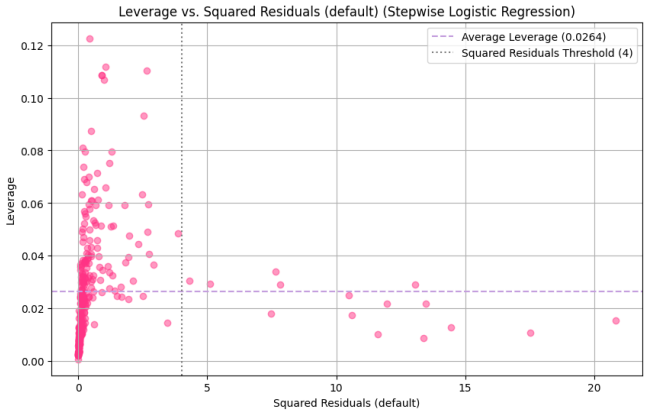

Figura 12: Residuales vs Leverage del modelo logístico ajustado.

Mencionamos finalmente que los valores de AIC y BIC del modelo seleccionado son 239.3125 y 269.0223, respectivamente. Comparando estos valores con los del modelo completo, el cual tiene un AIC de 239.436 y un BIC de 291.428, notamos que el AIC permanece muy similar con una muy pequeña mejora en el modelo más simple pero el BIC sí muestra una disminución considerable, lo cual indica que ambos modelos tienen un ajuste similar en términos de verosimilitud pero el modelo seleccionado logra una mejor compensación entre ajuste y complejidad, sin mostrar una pérdida significativa en su capacidad explicativa.

Por lo tanto, la concordancia obtenida entre ambos enfoques (con modelos bayesianos y con método stepwise) refuerza la robustez de los resultados obtenidos e indica que estas variables son las que realmente aportan información relevante para explicar la probabilidad de que se presente una enfermedad cardiovascular. Con esto, podemos proponer un modelo más simple que el completo analizado en el artículo original y con una sólida validez estadística, y sobre todo que sea capaz de capturar de manera eficiente las principales características determinantes del riesgo cardiaco dentro del conjunto de datos analizado.

4. Conclusiones

En el presente trabajo se analizaron y replicaron dos estudios que aplican modelos de regresión lineal y logística en contextos distintos. El primero en ecología, sobre el planeador de Leadbeater, y el segundo en salud, sobre predicción de enfermedades cardiacas. En ambos casos se buscó no solo reproducir los resultados originales, sino también proponer alternativas metodológicas que fortalecieran la interpretación y el desempeño de los modelos.

En el análisis del artículo “Relationship between body weight and elevation in Leadbeater’s possum (Gymnobelideus leadbeateri)” se comprobó que las variables de elevación y sexo son efectivamente significativas para explicar el peso del individuo, y que el modelo propuesto por el artículo cumple razonablemente los supuestos clásicos de la regresión lineal. Sin embargo, como demostramos en la regresión

agrupada por Caja Nido, el descarte de la variable de *NestBox* por los autores, decrece significativamente la posibilidad de obtener un mejor ajuste. Esto nos indica que el descarte de una variable por métodos empíricos no siempre es la mejor forma de descartar variables, y de hacerse este descarte, debe estar muy bien justificado y analizar las consecuencias de hacerlo. Un fenómeno interesante a notar es que la R^2 de la regresión *Elastic Net* es más pequeña que la de la regresión lineal; esto es normal pues la regresión lineal minimiza el error cuadrático sin ninguna forma de penalización mientras que el *Elastic Net* se enfoca sí, en minimzar el error cuadrático, pero también en reducir la varianza del modelo pues, al forzar algunos coeficientes hacia cero, el modelo se vuelve más estable ante pequeñas perturbaciones en los datos.

En cuanto al artículo "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy", se verificó que el modelo de regresión logística presenta un desempeño alto y estable, con métricas de precisión y exactitud similares a las reportadas por los autores originales. A través de las propuestas adicionales basadas en un modelos bayesianos y selección de variables mediante el método stepwise, se demostró que es posible obtener un modelo más parsimonioso, con igual capacidad predictiva pero mayor interpretabilidad, identificando como variables más relevantes a exang, ca, oldpeak, cp, thal, sex y thalach. Este resultado resalta la importancia de la selección de variables como herramienta esencial para obtener modelos más parsimoniosos e interpretables.

En conjunto, ambos estudios ilustran cómo las metodologías de regresión, tanto lineales como logísticas, constituyen herramientas fundamentales dentro de la ciencia de datos, no solo por su capacidad predictiva, sino también por su valor interpretativo. Además, evidencian la necesidad de aplicar procedimientos de validación y selección de variables que permitan construir modelos estadísticamente robustos.

Finalmente, este trabajo destaca la relevancia de replicar y contrastar estudios científicos mediante enfoques alternativos, pues esto fortalece la confianza en los resultados y genera valor para la creación de modelos cada vez más robustos.

Referencias

[Rimal et al., 2025] Rimal, Y., Sharma, N., Paudel, S., et al. (2025). Comparative analysis of heart disease prediction using logistic regression, svm, knn, and random forest with cross-validation for improved accuracy. *Scientific Reports*, 15:13444.

[Williams et al., 2021] Williams, J. L., Harley, D., Watchorn, D., McBurney, L., and Lindenmayer, D. B. (2021). Relationship between body weight and elevation in leadbeater's possum (gymnobelideus leadbeateri). *Australian Journal of Zoology*, 69:167–174.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.