

Ensayo reflexivo

Comparación de dos opciones analíticas en la predicción del rendimiento académico: árboles de clasificación *versus* regresión logística

Comparison of two analytical options in academic performance prediction: classification trees *versus* logistic regression

Jorge Bacallao Gallestey¹ <https://orcid.org/0000-0001-7961-9970>

Raúl Fernández Regalado^{2*} <https://orcid.org/0000-0001-7407-9135>

Lázaro Elier Alba Zayas² <https://orcid.org/0000-0003-3822-7799>

¹Centro de Investigaciones y Referencia de Aterosclerosis de La Habana. Cuba.

²Universidad de Ciencias Médicas de La Habana. Instituto de Ciencias Básicas y Preclínicas "Victoria de Girón" - Facultad de Ciencias Médicas. La Habana. Cuba.

*Autor para la correspondencia: raulfernandezregalado2023@gmail.com

RESUMEN

Los árboles de clasificación y regresión y la regresión logística son dos técnicas estadísticas utilizadas frecuentemente en la investigación biomédica, las cuales están contenidas en varios programas estadísticos. Sin embargo, no son muy conocidas sus ventajas y desventajas y las condiciones para la aplicación de cada una, de ahí que el propósito de este artículo sea comparar, desde una perspectiva didáctica, los dos recursos analíticos en el contexto de los estudios de predicción. El texto enfatiza las ventajas y desventajas relativas de esas técnicas y en sus valores y usos complementarios. Se procura minimizar el empleo del lenguaje técnico y en su lugar se exponen argumentos muy permeables a la intuición y al

razonamiento práctico. Ambas opciones se ilustran con los resultados de una investigación relativa al pronóstico del rendimiento académico en estudiantes del primer curso de la carrera de medicina en la Facultad de Ciencias Médicas “Victoria de Girón”, de la Universidad de Ciencias Médicas de La Habana.

Palabras clave: árboles de clasificación y regresión; regresión logística; rendimiento académico; estudios de predicción.

ABSTRACT

Classification and regression trees and logistic regression are two statistical techniques frequently used in biomedical research, which are contained in several statistical programs. However, their advantages and disadvantages, and the conditions for the application of each, are not well known. The purpose of this article is to compare, from a didactic perspective, the two analytical resources in the context of prediction studies. This paper emphasizes the advantages and relative disadvantages of these two techniques and their values and complementary uses. Efforts are made to minimize the use of technical language, and instead, highly intuitive and practical reasoning arguments are presented. Both options are illustrated with the results of an investigation related to the prognosis of academic performance in first-year medical students at the Faculty of Medical Sciences “Victoria de Girón”, University of Medical Sciences of Havana.

Keywords: classification and regression trees; logistic regression; academic performance; prediction studies.

Recibido: 28/04/2025

Aprobado: 07/02/2026

Introducción

Los árboles de clasificación y regresión (CART, del inglés *classification and regression trees*)⁽¹⁻⁴⁾ constituyen un procedimiento de naturaleza inductiva, cuyas condiciones óptimas de aplicación son las fases exploratorias de una investigación en las que aún es prematuro intentar poner a prueba hipótesis o conjeturas con suficiente respaldo empírico y/o teórico. Cuando se dispone de una gran cantidad de casos, los CART operan como un recurso que recuerda los caminos algorítmicos de la inteligencia artificial.

En este artículo, además de analizar las características de los CART, se presentan los atributos de la regresión logística (RLO) que marcan un contraste con los primeros, con la intención de comparar, desde una perspectiva didáctica, ambos recursos analíticos en el contexto de los estudios de predicción. Al final del acápite Desarrollo se ilustran las dos técnicas estadísticas con los resultados de un trabajo sobre el pronóstico del rendimiento académico en estudiantes de primer año de medicina.

Desarrollo

Los actores principales en un CART son usualmente una variable dependiente binaria y un conjunto de variables medidas en cualquier escala (desde cualitativas hasta cuantitativas continua), que operan como posibles variables explicativas. Esta plasticidad en la naturaleza de las variables es una de sus obvias ventajas. ¿Qué entrega el CART? ¿Cuál es su principal producto? Una partición de todos los casos en subconjuntos (nodos en la terminología usual de los CART) disjuntos, que reproducen la clasificación inicial en la dicotomía representada por la variable dependiente.

Para ilustrarlo con un ejemplo simple: en un gran conjunto de pacientes que han recibido un trasplante renal, un cierto porcentaje sobrevive al término de los 5 años posteriores al trasplante. Esa variable binaria, que identifica al paciente como fallecido o sobreviviente, funciona como variable dependiente en el esquema racional del CART, a la que se añade un número teóricamente ilimitado de variables independientes que operarían como posibles predictores de la condición al cabo de los 5 años de seguimiento.

El derrotero analítico de un CART es la construcción de varios nodos llamados “filiales” a partir de un nodo inicial llamado “parental”, una mnemotecnia retórica que alude al crecimiento de un árbol en diferentes ramas a partir de la raíz.

Aunque existen varios algoritmos de optimización para construir los nodos filiales a partir de los parentales, en este artículo nos referimos al que se basa en maximizar la “limpieza” del nodo filial, lo cual entenderemos como la posibilidad de clasificar a los pacientes con el mínimo error posible.⁽⁵⁾

Si usamos de nuevo el símil del trasplante renal, supongamos que todos los pacientes que han sobrevivido al cabo de los 5 años recibieron el órgano de un paciente vivo, y para mayor precisión, de un familiar con un alto grado de compatibilidad en los distintos sistemas que definen esta condición. Y, en el otro extremo, todos los fallecidos recibieron el órgano de un donante muerto. El tipo de donante definiría entonces una partición en la cual la clasificación se realizaría sin ningún error. El esquema siguiente (Figura 1), ficticio y virtualmente imposible, ilustra la idea anterior, para mejor comprensión del lector.

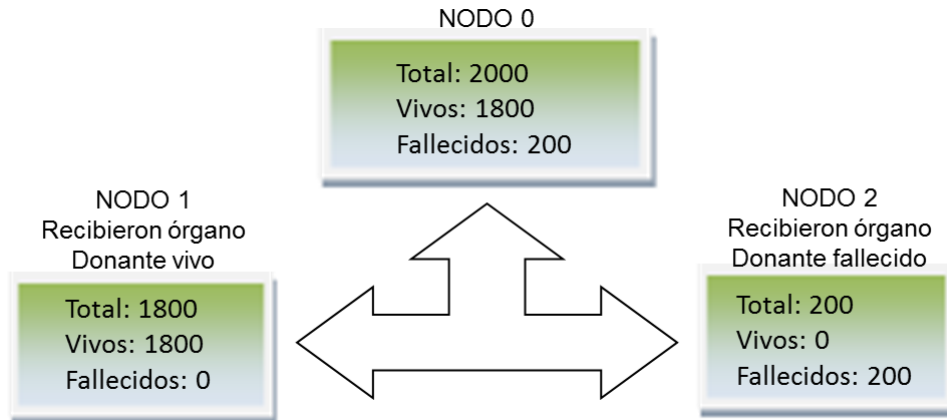


Fig. 1- Ejemplo sobre supervivencia a un trasplante renal según tipo de donante.

Es obvio que los nodos del ejemplo ficticio anterior son nodos “terminales”, a partir de los cuales no puede “crecer” ninguna nueva rama del árbol. El conjunto de todos los nodos terminales define la partición que constituye el producto básico y final del CART.

Lo habitual es que el resto de las variables explicativas den lugar a nuevas ramificaciones del árbol, con el criterio –y éste es uno de los aportes más relevantes de los CART–, de maximizar la “limpieza” del nodo, lo cual, en el caso de las variables continuas, implica encontrar el punto de corte óptimo para variables de ese tipo, o las opciones discretas para las variables de naturaleza discontinua, ya se trate de variables cuantitativas o cualitativas. En los softwares disponibles, el investigador controla el número mínimo de casos que deben contener los nodos parentales y filiales. Un nodo con pocos casos es poco informativo. Esta circunstancia explica claramente por qué la utilidad práctica de los CART depende, y se incrementa, con el número de casos.

Uno de los principales atractivos de los CART, que lo distinguen, como veremos más adelante, de la otra opción metodológica de este texto, es que constituyen un

procedimiento no paramétrico, o en otros términos, que no requiere supuestos de ningún tipo acerca de la distribución probabilística de las variables, ni de relaciones funcionales específicas como la linealidad o la monotonía de las relaciones entre variables, que sí son clave para otros recursos. La versión que consideramos en este artículo, basada en la maximización de la limpieza, no descansa en ningún proceso inferencial que vaya más allá de cálculos descriptivos simples. Una variable que define la ramificación en nodos filiales puede “aparecer” más de una vez en la definición de dichos nodos, hacia el logro de la partición óptima. Esta característica prácticamente singulariza los CART como recurso para la identificación de variables relevantes.

Los fundamentos y los usos de la RLO se exponen en muchos textos de estadística básica y aplicada que el lector puede consultar.^(6,7)

Una nota cautelar bien importante es que la RLO puede utilizarse con dos fines, con requerimientos diferentes: el fin predictivo (en el que se centra el ejemplo de este texto) y el fin explicativo. Un error muy común consiste en pretender combinar ambas aplicaciones en un único ejercicio práctico.

Con fines predictivos, en el escenario que nos ocupa con los mismos protagonistas que se describen en el apartado anterior, la meta principal de la RLO es contar con el cálculo de una probabilidad final: la de pertenecer a cualquiera de las dos categorías de la variable dependiente. Vale aclarar que nos referimos a la regresión logística binaria, que supone una variable dependiente dicotómica, y no a otras modalidades, como la RLO politómica, por lo que estas probabilidades estimadas son complementarias respecto a la unidad. Una vez calculada, o más precisamente, estimada esa probabilidad, es elemental evaluar sus propiedades métricas, como son la sensibilidad, la especificidad y los valores predictivos positivos y negativos, lo cual suele hacerse mediante la aplicación de otros recursos analíticos como las llamadas curvas ROC.⁽⁸⁾

Cuando la RLO se aplica con fines predictivos, la selección de los predictores puede hacerse bajo criterios laxos basados en conocimientos teóricos y/o empíricos, y bajo consideraciones de costo-beneficio. La ganancia que se consiga en cuanto a capacidad predictiva debe contrastarse con el costo que implique poder medir la variable en cuestión y con las posibles implicaciones éticas de esa decisión. No hay otras provisiones que observar.

Todo lo contrario ocurre cuando se aplica con fines explicativos, lo que supone un veredicto sobre la importancia de la variable en cuestión en relación con la variable dependiente. Este uso de la RLO descansa en la aplicación de recursos propios de la estadística inferencial, que se basan en supuestos sobre la distribución de las variables, indispensables, tanto para estimar los parámetros del modelo como para poner a prueba hipótesis acerca de su relevancia.

En esta aplicación con fines explicativos, es tan frecuente como erróneo elegir las variables que se ingresan al modelo entre aquellas que resultaron significativas, en el sentido estadístico, en pruebas de hipótesis univariadas previas sobre su asociación con la variable dependiente. Más allá del descrédito de las pruebas de significación como recurso inferencial,^(9,10) el error descansa en el fundamento mismo de la RLO, que consiste en fijar el valor de todos los predictores menos uno, y en constatar cuánto influyen las variaciones de este único predictor sobre la variable dependiente. Es fácil advertir que esta práctica ignora la covariación entre los predictores, y que únicamente sería aceptable cuando se disponga de predictores mutuamente independientes, lo cual es completamente irreal.

La aplicación con fines explicativos exige una selección de las variables a partir de consideraciones teóricas sobre los roles de las variables, que suelen tener patrones de asociación complejos, que incluyen la presencia de confusores (variables que son causa común de la presunta causa y el presunto efecto), colisores (variables que son efecto común de la presunta causa y del presunto efecto) y mediadores

(que son variables en la trayectoria causal que liga a la presunta causa con el presunto efecto).⁽¹¹⁾

La predicción del rendimiento académico: ejemplo con resultados y comentarios

En el ejemplo que se presenta y discute a continuación, el objetivo fundamental fue identificar predictores relevantes del rendimiento académico de los estudiantes de primer año de la carrera de medicina, y poner a prueba su capacidad predictiva. Es conveniente resaltar que antes de aplicar la encuesta y las pruebas psicológicas, los estudiantes fueron informados de las características y propósitos de la investigación, el carácter voluntario y anónimo de su participación y el compromiso de los investigadores de mantener la confidencialidad de los datos y usarlos solamente con fines académicos. El proyecto de esta investigación fue aprobado por el Comité de Ética de la Investigación del ICBP “Victoria de Girón”.

La población estuvo constituida por todos los estudiantes de nuevo ingreso a la carrera de medicina del curso 2022, de los cuales se seleccionó una muestra mediante muestreo aleatorio estratificado por grupo docente. A todos los estudiantes incluidos en el estudio se les aplicó una encuesta diseñada por los autores y validada por el criterio de cuatro profesores, sin conflictos de interés y con experiencia en el área docente (3 Profesores Titulares y 1 Profesora Auxiliar), a quienes se solicitó que emitieran sus opiniones sobre el contenido y la forma del instrumento mediante un cuestionario de seis preguntas. Como resultado del proceso de validación, tres de los consultados realizaron sugerencias de carácter formal que se tuvieron en cuenta para la versión definitiva que se aplicó. Se realizó además un ensayo piloto con un pequeño grupo de estudiantes no seleccionados en la muestra, quienes dieron fe de la comprensibilidad de las preguntas e instrucciones.

En adición a la encuesta referida y para evaluar el rendimiento intelectual, se aplicó a cada estudiante el test de dominó.⁽¹²⁾

El nivel de desarrollo de motivación profesional y otros rasgos de la personalidad fueron explorados mediante la “técnica de los diez deseos”, la prueba de Rotter y el IPJ (inventario de problemas juveniles). Se utilizó asimismo el método clínico para analizar las expresiones conscientes de los alumnos en esas pruebas.⁽¹³⁾

La variable dependiente o a predecir: rendimiento académico, consistió en una evaluación dicotómica en las categorías “satisfactorio” y “no satisfactorio”, según los criterios que se describen a continuación.

- Satisfactorio: el estudiante aprobó todas las asignaturas de ciencias básicas biomédicas en su primera convocatoria, con independencia de la calidad de la nota obtenida en cada una de ellas.
- No satisfactorio: el estudiante desaprobó al menos una asignatura de ciencias básicas biomédicas en la primera convocatoria.

Los predictores considerados fueron:

- *El escalafón.* Resume la trayectoria académica del estudiante al ingreso; históricamente ha resultado relevante. Esta variable se definió de la siguiente manera: promedio de las notas del estudiante en el preuniversitario y promedio de las pruebas de ingreso a la universidad (matemática, historia y español). El escalafón fue la media de ambos promedios.
- *La motivación.* Inicialmente fue clasificada en extrínseca, intrínseca o ausencia de motivación, según los resultados de las pruebas para su medición. Posteriormente fue convertida a una escala binaria de presencia-ausencia.
- *Las horas de estudio semanales.* Fueron las reportadas en la encuesta por los mismos estudiantes.
- *El sexo.* Se incluyeron las categorías habituales de masculino y femenino, reportadas en la encuesta por cada estudiante.

Se obtuvieron dos árboles de clasificación que se exponen y explican a continuación.

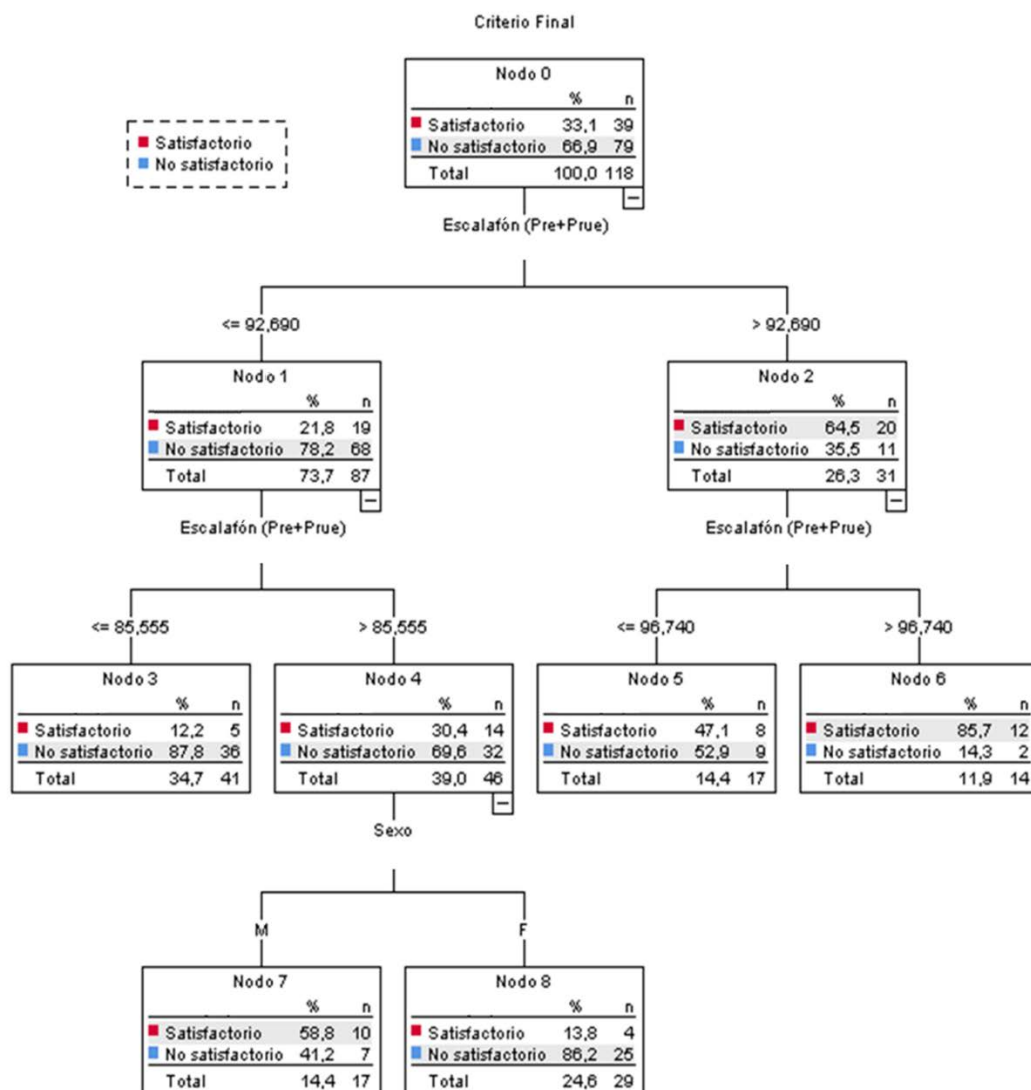


Fig. 2- Árbol de clasificación en el que se aprecia la participación de la variable escalafón en la definición de todos los nodos terminales.

En la Figura 2 se advierte de inmediato que todo el proceso está dominado por la variable escalafón, que interviene en la definición de todos los nodos terminales, con diferentes puntos de corte. La primera partición se origina en la selección del punto de corte 92,69. Treinta y uno de los 118 estudiantes incluidos en el estudio se encontraban por encima de este punto de corte, y entre ellos el porcentaje de

casos “no satisfactorios” se redujo de 66,9 a 35,5. El examen de los nodos 3, 4, 5 y 6 permite advertir de inmediato la relación monótona entre el rendimiento y la historia académica descrita por el escalafón. En efecto, el porcentaje de casos “no satisfactorios” disminuye a medida que el estudiante rebasa puntos de corte más exigentes en su escalafón: de 87,8 a 69,6, a 52,9 y finalmente a 14,3.

Es interesante advertir el rol del sexo en la definición de los nodos filiales 7 y 8, que emergen del nodo 4. El porcentaje de casos “no satisfactorios” entre las muchachas (86,2) es más del doble que entre los varones (41,2).

Este resultado, un tanto sorprendente, motivó a los autores a “forzar” la entrada de la variable sexo en la resolución del árbol. “Forzar” la entrada de una variable es una de las opciones en el manejo práctico de los CART, que se aplica cuando se considera conveniente utilizar algún criterio de estratificación. Esa opción se aplicó en el árbol siguiente (Figura 3), cuya interpretación se deja ya a cargo del lector.

No obstante, es bien llamativo el hecho de que para varones y muchachas emergen puntos de corte óptimos, notablemente diferentes. Mientras que los primeros tienen un buen pronóstico (más de 60 % de resultados “satisfactorios”) con un punto de corte de 85,49, el buen pronóstico en las muchachas requiere un punto de corte superior en más de 9 puntos (93,59).

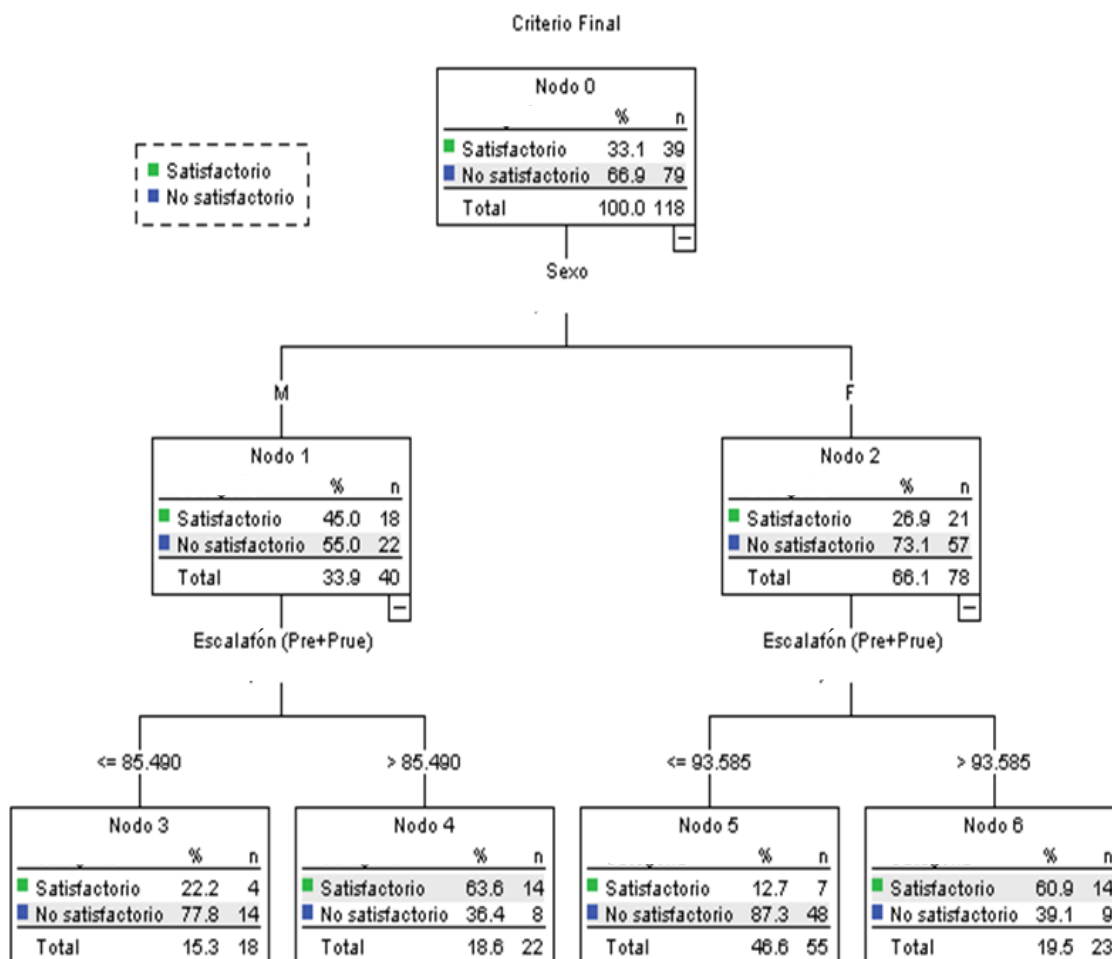


Fig. 3- Árbol de clasificación obtenido al “forzar” la entrada de la variable sexo.

Veamos ahora los resultados básicos que arrojó la RLO, para cuya aplicación se resolvió incluir los cuatro predictores y no sólo los sugeridos por el CART.

En primer lugar, se reproduce el componente esencial que incluye el estadístico de Wald y su valor de “p” asociado (Tabla 1).^(5,14)

Tabla 1- Estadístico Wald y valor de “p” asociado para cada predictor en la RLO

Predictor	Wald	“p”
Sexo	6,255	0,012
Horas de estudio semanales	3,541	0,060
Motivación	0,695	0,404
Escalafón (Preuniversitario + Pruebas de ingreso)	12,747	0,000

En los valores de “p” se advierte completa coincidencia con el resultado que arrojó el CART, con el escalafón y el sexo en ese orden de relevancia.

Finalmente, se reproduce la importante matriz de clasificación (Tabla 2), que emerge luego de estimar mediante el modelo de la RLO la probabilidad de un resultado “satisfactorio” o “no satisfactorio”.

Tabla 2- Matriz de clasificación obtenida a partir de la estimación mediante RLO de la probabilidad de un resultado “satisfactorio” o “no satisfactorio”

Observado		Predicho		
		Resultado		Porcentaje correcto
		Satisfactorio	No satisfactorio	
Resultado	Satisfactorio	19	13	59,4
	No satisfactorio	7	67	90,5
Porcentaje total				81,1

El porcentaje de clasificación correcta (la sensibilidad) entre los “no satisfactorios” fue de 90,5 %, y entre los “satisfactorios” de sólo 59,4 % (la especificidad). Estos porcentajes dependen, por supuesto, del punto de corte que el investigador elija para la probabilidad estimada, según la importancia que atribuya a la sensibilidad y a la especificidad que, como es bien conocido, son atributos antagónicos. La alta sensibilidad responde a una elección baja del punto de corte para la probabilidad de “fracaso”, que permita adoptar medidas oportunas con los estudiantes con un pronóstico adverso.

En el apartado siguiente se resumen las ventajas y desventajas relativas de ambos recursos analíticos.

Consideraciones finales

Los CART son preferibles para identificar variables relevantes, porque a esos fines no requieren supuestos sobre la distribución de las variables y por consiguiente operan sobre una base no paramétrica. Por contraste, con esos propósitos la RLO es un recurso paramétrico que exige supuestos bastante restrictivos, no sólo sobre la distribución probabilística de las variables, sino sobre su patrón de asociación. Si convencionalmente llamamos “éxito” y “fracaso” a los dos valores posibles de la variable dependiente, la relación entre un predictor cualquiera y la probabilidad de “éxito” (o de “fracaso”) debe ser monótona (creciente o decreciente) y lineal.

Las ventajas de los CART se atenúan si hay pocos casos. La RLO es menos exigente respecto del tamaño de la muestra. Los CART aportan, como parte de sus salidas básicas, puntos de corte óptimos para los predictores de naturaleza continua o cuantitativa discreta.

Los CART prácticamente se autoexplican, por lo que son más fáciles de interpretar. Sus resultados fundamentales se expresan mediante frecuencias relativas y absolutas al alcance de cualquier lector.

Más allá de sus diferencias y ventajas y desventajas relativas, CART y RLO tienen valores complementarios. Bajo las condiciones que habitualmente pueden encontrarse en ensayos clínicos controlados, o en estudios observacionales con diseños prospectivos adecuados, o con diseños transversales en los que sea posible establecer de modo inequívoco la precedencia cronológica de las variables, los CART pueden utilizarse como un paso inicial para identificar predictores relevantes y la RLO, para efectuar la predicción mediante la estimación de las probabilidades.

Referencias bibliográficas

1. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression Trees. Monterey, CA: Brooks/Cole Publishing; 1984.
2. Loh W-Y. Classification and regression trees. WIREs: Data Mining Knowl Discov [Internet]. 2011 [citado 30 Ene 2025];1(1):14-23. Disponible en: <https://wires.onlibrary.wiley.com//doi/10.1002/widm.8>
3. Carrasco RA, Bueno I, Montero JM. Árboles de clasificación y regresión. En: Fernández Avilés G, Montero JM, editores. Fundamentos de ciencia de datos en R [Internet]. Madrid: McGraw-Hill Interamericana de España S. L.; 2024 [citado 20 Mar 2025]. Disponible en: <https://cdr-book.github.io/cap-arboles.html>
4. Carrizosa E, Molero-Río C, Romero Morales D. Mathematical optimization in classification and regression trees. TOP [Internet]. 2021 [citado 20 Mar 2025];29:5-33. Disponible en: <https://link.springer.com/content/pdf/10.1007/s11750-021-00594-1.pdf>
5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Nueva York, NY: Springer; 2009.
6. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. Nueva York, NY: Springer; 2021.
7. Martínez Pérez JA, Pérez Martín PS. Regresión logística. Medicina de Familia. SEMERGEN [Internet]. 2024 [citado 21 Mar 2025];50(1):1-7. Disponible en: <https://www.elsevier.es/es-revista-familia-semergen-40-pdf-S1135389323001661>
8. Schalltmann P. Statistics in diagnostic medicine. Clin Chem Lab Med [Internet]. 2022 Mar 31 [citado 23 Feb 2025];60(6):801-807. Disponible en: <https://degruyterbrill.com/document/doi/10.1515/cclm-2022-0225/html>
9. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat [Internet]. 2019 [citado 30 Ene 2025];73(Suppl 1):S235-245. Disponible en:

<https://tandfonline.com/doi/full/10.1080/00031305.2018.1527253?scroll=top&needAccess>

10. Hurlbert SH, Levine RA, Utts J. Coup de Grâce for a Tough Old Bull: “Statistically Significant” Expires. Am Stat [Internet]. 2019 [citado 30 Ene 2025];73(suppl 1):S352-357. Disponible en:
<https://tandfonline.com/doi/full/10.1080/00031305.2018.1543616?scroll=top&needAccess>
11. Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. Eur J Epidemiol [Internet]. 2017 [citado 16 Feb 2025];32(1):3-20. Disponible en:
<https://link.springer.com/article/10.1007/s10654-017-0230-6>
12. Anstey E. D-48 Tests de Dominós. Manual. 12 ed. Madrid: TEA Ediciones; 1999.
13. Martínez Angulo MR. Manual de técnicas de exploración psicológica. La Habana: Editorial Pueblo y Educación; 2013.
14. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B: Statistical Methodology [Internet]. 2005 [citado 2 Mar 2025];67(2):301-320. Disponible en:
<https://studylib.net/doc/8388343/regularization-and-variable-selection-via-the-elastic-net>

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses.