Árboles de clasificación vs regresión logística en el desarrollo de competencias genéricas en ingeniería

Jorge Pérez-Rave[1], Favián González Echavarría[2]

[1] IDINNOV S.A.S, Grupo de investigación IDINNOV,
Colombia

[2] Universidad de Antioquia, Departamento de Ingeniería Industrial,
Facultad de Ingeniería,
Colombia

investigacion@idinnov.com, favian.gonzalez@udea.edu.co

Resumen. Desde un enfoque experimental, el desempeño de Regresión Logística vs Árboles de Clasificación es evaluado en el contexto de dos competencias genéricas en ingeniería (razonamiento cuantitativo y comprensión lectora). Dichos métodos incorporan dos escenarios de predictores por separado (solo indicadores y solo constructos; estos últimos, derivados de Análisis de Componentes Principales: ACP). La muestra es de 7,395 instancias de Saber Pro 2015-3, 2014-3 (Colombia). El estudio considera: entrenamiento (70 % de la muestra), predicción (30 % restante) y experimentación (176 observaciones originales; diseño de tres factores: método, tipo de predictor y competencia). La variable respuesta es una nueva métrica (Tasa Subyacente de Aciertos obtenida por ACP). Ambos métodos presentan similar desempeño en el escenario de solo indicadores, pero no en el escenario de constructos (Regresión Logística, mejor desempeño).

Palabras clave. Regresión logística, árbol de clasificación, competencias en ingeniería.

Classification Trees vs. Logistics Regression in the Generic skill Development in Engineering

Abstract. From an experimental approach, the performance of Logistic Regression vs Classification Trees is evaluated in the context of two generic engineering skills (quantitative reasoning and reading comprehension). These methods incorporate two separate predictor scenarios (indicators only, and only constructs derived from Principal Component Analysis: ACP). The sample is 7,395 instances of Saber Pro 2015-3, 2014-3 (Colombia). The study considers: training (70% of the sample), prediction (30% remaining) and experimentation (176 original observations, design of three factors: method, type of predictor and competence). The response variable is a new metric (Underlying Asset Ratio obtained by ACP). Both methods present similar performance in the scenario of only indicators, but not in the scenario of constructs (Regression Logistics, better performance).

Keywords. Logistic regression, classification tree, engineering skills.

1. Introducción

El grado de desarrollo de competencias en los estudiantes de educación superior, a la larga influencia los resultados de las empresas que les contraten, así como la calidad de vida de sus familias y de la sociedad en general.

Es un tópico de creciente interés para gobiernos, academia y organizaciones en general, los cuales llaman constantemente a comprender sus determinantes para inducir políticas de mejora en los procesos formativos (Bahamón y Reyes, 2014). Sin embargo, los antecedentes del desarrollo de competencias en el estudiante es un tema complejo y multifactorial (Vargas, 2007), donde la mayoría de factores están por descubrir.

En ellos hay involucradas dimensiones del ser, el saber y el hacer (generales y específicas) que se forjan con el propio estudiante, su familia, escuela, institución de educación superior y el resto de entorno (Fernández, 2011).

Entender dichos antecedentes amerita, primero, concertar la forma de medir tal constructo (factor latente). Una posibilidad, comúnmente empleada por los gobiernos, son las pruebas estandarizadas de estado. En Colombia se utiliza la prueba Saber Pro, administrada por el Instituto Colombiano para la Evaluación de la Educación (ICFES). Esta evalúa competencias genéricas y específicas (propias de cada programa académico) cuando el estudiante ha cursado, como mínimo, el 75 % de los créditos del programa (Caicedo, Guerrero & López, 2016).

Hoy día, los avances derivados de utilizar bases de datos de Saber Pro se han centrado en describir o segmentar resultados (Cañón & Jiménez, 2017; Bahamón y Reyes, 2014; Delgado, 2013), relacionar puntuaciones (León, Amaya y Orozco, 2012), crear nuevos índices (Caicedo, Guerrero, & López, 2016) o explorar determinantes (Cortés y Piñeros, 2015; Torres, Vélez & Altamar, 2015; Ramírez, 2014; Castellanos, Mojica & Rivera, 2014; Gil et al., 2013). No obstante, la mayoría de trabajos centrados en modelos ha empleado la regresión clásica, lo cual abre posibilidades para explorar otros métodos, como los de clasificación.

En adición, los modelos propuestos han tendido a incorporar las variables regresoras tal como provienen de la base de datos de Saber Pro, dejando oportunidades para probar constructos (grupos de indicadores). Esto resulta útil, ya que el desarrollo de competencias en el estudiante es un fenómeno de campos sociales y humanos, que desde la teoría es explicado por medio de conceptos complejos (Ej.: condiciones socioeconómicas del estudiante y su familia,).

Para medir tales conceptos es necesario convertirlos en constructos que subyacen en grupos de indicadores observables, altamente correlacionados. De este modo, no solo se aprovecha la información de cada indicador por separado, sino también de grupos de indicadores latentes (abstracciones), propias de ciencias sociales y humanas.

Entonces, al emplear constructos en modelos paramétricos (Ej: regresiones) o no paramétricos (Ej: basados en árboles), puede evaluarse si mejora el desempeño y/o la interpretación de los modelos actuales. En consecuencia, se obtendría nuevos hallazgos para favorecer la toma de decisiones institucionales y estatales.

Con miras a explorar el tema, el objetivo es evaluar el desempeño de dos tipos de modelos de clasificación (Regresión Logística y Árboles de Clasificación), probándolos solo con indicadores, y, por aparte, solo con constructos, sobre dos competencias genéricas de Saber Pro en ingeniería (Razonamiento cuantitativo y Comprensión lectora).

Este trabajo realiza tres contribuciones. La primera es que aporta un nuevo enfoque para el uso de los datos de las pruebas Saber Pro, abordando los resultados de estas a través métodos de clasificación (con y sin constructos).

La segunda es que provee evidencia empírica sobre la comparativa entre Regresión Logística y Árboles de Clasificación. Esto último resulta pertinente, debido a la ausencia de consenso sobre qué método presenta mejor desempeño; más aún, las interpretaciones tienden a excusarse en el tipo de problema. Por un lado, hay quienes argumentan a favor de los métodos clásicos (Ej: Regresión Logística), destacando mejor capacidad de predicción en muestras no-entrenamiento, entre otras bondades (Fernández et al., 2016; James et al., 2015); en cambio, otros autores concluyen lo contrario (Zayeri et al., 2016; Coussement, Van den Bossche & De Bock, 2014).

No obstante, los casos de estudio y sus contextos no necesariamente son comparables, por lo que incorporar factores contingentes dentro de una misma comparativa y distinguir sus efectos, puede aportar nuevas explicaciones.

En ese sentido, este trabajo se diferencia de comparativas clásicas, basadas en curvas ROC con AUC, al emplear un diseño experimental de tres factores fijos: tipo de método (Regresión Logística, Árboles de Clasificación), tipo de predictor (Indicadores, Constructos) y tipo de competencia (Comprensión Lectora, Razonamiento Cuantitativo). La variable respuesta es el desempeño de los métodos, en términos de la Tasa Subyacente de Aciertos.

La elección de las competencias se fundamentó en incorporar dos tipos de habilidades, “dura” (razonamiento cuantitativo) y otra “blanda” (comprensión lectora). Además, se busca separar el efecto del “tipo de competencia”, ya que el problema específico de prueba ha

tendido a ser la explicación más común ante eventuales controversias.

La tercera contribución se debe a que las comparativas entre métodos de clasificación han tendido a realizarse considerando, individualmente, los índices de la matriz de confusión, los cuales pueden arrojar resultados en conflicto. Cuando esto se da, el analista tiende a inclinarse por concluir de forma individual para cada métrica, o recurre a estrategias no necesariamente reproducibles u objetivas para fijar una posición global. Este estudio busca evitar dicha responsabilidad del analista, utilizando varios índices de la matriz de confusión como indicadores proxy de una nueva métrica: Tasa Subyacente de Aciertos. Para ello, se adopta un enfoque multivariado, basado en Análisis de Componentes Principales (ACP).

A nivel específico, se busca proveer respuestas a cuatro interrogantes, en el contexto de competencias genéricas de Saber Pro en ingeniería:

P.1 ¿El uso exclusivo de constructos en los métodos de clasificación (Regresión Logística y Árboles de Clasificación) mejora la predicción de estudiantes con alto/bajo desarrollo de competencias?

P.2 ¿Qué método, entre Regresión Logística y Árboles de Clasificación, presenta mejor desempeño respecto a la predicción de estudiantes con alto/bajo desarrollo de competencias?

P.3 ¿Cuáles son los principales predictores (indicadores y constructos) del alto/bajo desarrollo de competencias?

P.4 ¿Los principales predictores (indicadores y constructos) del alto/bajo desarrollo de competencias genéricas de Saber Pro en ingeniería, difieren según el tipo de competencia (Razonamiento cuantitativo, Comprensión lectora)?

La sección 2 aporta el referencial teórico. La sección 3 provee el procedimiento propuesto. La sección 4 ofrece los resultados obtenidos y la sección 5 condensa la discusión.

2. Referencial teórico

2.1. Estudios recientes sobre Saber Pro

Bahamón & Reyes (2014) caracterizan estudiantes con alto y bajo rendimiento en las pruebas Saber Pro, a la luz de capacidad intelectual, elementos sociodemográficos y académicos. La muestra fue de 68 estudiantes de psicología que presentaron pruebas en el año 2012. Usaron T-Student. Concluyen que aquellos estudiantes de mayor capacidad intelectual suelen lograr puntajes más altos en las pruebas. El estrato no resultó ser un elemento diferenciador. Aquellos estudiantes provenientes de colegios privados presentaron mejor rendimiento en las pruebas. Los estudiantes con puntajes más altos suelen tener orientación profesional hacia actividades lingüísticas, humanitarias y de artes plásticas. Los estudiantes con mejor desempeño mostraron tener preferencia por técnicas de estudio progresivas y ordenadas (tomar notas, repasar diariamente, hacer esquemas…).

León, Amaya & Orozco (2012) estudian si existe relación entre la habilidad de comprensión lectora, la inteligencia y el desempeño de estudiantes de Psicología en las pruebas Saber Pro 2011. La muestra fue de 45 estudiantes. Se usó correlación de Spearman. Se observó una correlación positiva-fuerte entre el desempeño en las pruebas y los siguientes índices: de memoria general y de inteligencia general. El desarrollo estratégico de la comprensión lectora a través del fortalecimiento de la memoria y retentiva es un aspecto clave que influye en la calidad de los resultados obtenidos.

Caicedo, Guerrero & López, (2016) construyen un índice para la medición del nivel socioeconómico de los estudiantes de educación superior en Colombia, con el fin generar un insumo que apoye los procesos de evaluación y caracterización de los estudiantes. Tomaron la población de estudiantes que presentaron el examen Saber Pro durante el año 2012 (245461 estudiantes). Se hizo uso del cuestionario sociodemográfico de Saber Pro, así como el Análisis de Competentes Principales. El índice socioeconómico estuvo compuesto por: potencial económico del hogar (ingreso familiar, valor de la matrícula y nivel sisbén), capital cultural (nivel de

educación de la madre y el padre, y ocupación de la madre y el padre) y dotación de la vivienda (servicio de internet, televisión,…). Señalan que entre mayores los puntajes logrados en las dimensiones del índice socioeconómico, mejores fueron los resultados obtenidos en Saber Pro.

Cortés & Piñeros (2015) buscan identificar las variables que más influyen en el desempeño académico de los estudiantes universitarios de instrumentación quirúrgica en las pruebas Saber Pro 2010 y 2011. La muestra fue de 1116 estudiantes. Técnicas: Estadísticas descriptivas y ANOVA para estudiar diferencias en los resultados de las pruebas, según: edad, género, matricula pagada por padres, tenencia de computador, estrato, entre otros. Concluyen que en lectura crítica solo fue significativa la edad (relación positiva). En comunicación escrita, los mejores resultados se presentaron para género (mujeres), matricula pagada por los padres y tenencia de computador.

En inglés: estrato alto, acceso a internet desde casa y alto nivel de educación de padre y madre. En razonamiento cuantitativo solo fue significativa la matrícula pagada por padres.

Romero, Villarreal & Velandia (2015) estudian si existe diferencias en el desempeño en las pruebas Saber Pro entre dos universidades colombianas, una de modalidad virtual y otra presencial.

La muestra fue de 194 estudiantes de administración de empresas (136 en presencial y 58 en virtual) que presentaron las pruebas en el año 2012. Técnicas: estadística descriptiva, correlación de Spearman, prueba no paramétrica U de Mann-Whitney, ANOVA y Kruskall-Wallis. Concluyen que el desempeño en las pruebas Saber Pro en estudiantes de modalidad virtual tienden a ser más bajas en razonamiento cuantitativo, aunque la diferencia no fue significativa. Para lectura crítica y competencia ciudadana no se encontraron diferencias significativas. En conclusión, la modalidad no parece ser un factor diferenciador del desempeño.

2.1. Estudios sobre árbol de clasificación vs. regresión logística

Con base en literatura de referencia (ver Tabla 13) puede concluirse:

1) La mayoría de aplicaciones se han llevado en contextos de la salud (7 de 15).

2) Los tamaños de muestra son relativamente pequeños (mín.: 205, mediana: 460, Q3: 1245; con un valor aparentemente atípico de 39710).

3) Prácticamente la mitad de los estudios ha incorporado, dentro de la comparativa, al menos otro método, adicional a Árboles de Clasificación y Regresión Logística (7 de 15). 4)

En la mayoría de casos Árboles de Clasificación han mostrado mejor desempeño que Regresión logística (en 10 casos fue superior y en otro más se consideró empate), pero es de anotar que los problemas de prueba y contextos no necesariamente son comparables.

3. Materiales y métodos

3.1. Depuración de la base de datos

La selección de observaciones partió de un marco de 534.575 estudiantes que presentaron las pruebas Saber Pro en dos cohortes: 2014-3 (209.726) y 2015-3 (324.849). Los nombres, tal cual como provienen de la fuente ftp del ICFES, son: “SBPRO-20143-RGSTRO-CLFCCN-U_GEN” y SBPRO-20153-RGSTRO-CLFCCN-U_GEN, respectivamente. Se controlaron diversas características en el conjunto de datos (Ej.: municipio de residencia del estudiante, programas de ingeniería).

Estas se precisan en el campo “criterio” del Tabla 14. Siguiendo el paso a paso allí expuesto, es posible reproducir la preparación de la base de datos de instancias de prueba. En total se seleccionaron 7395 estudiantes, que equivale al 1,38 % del total alojado en las bases de datos de Saber Pro del ICFES (20143, 20153).

Esta cantidad de instancias no solo supera considerablemente el tercer cuartil del tamaño de muestra que han empleado estudios previos (Q3: 1.245; véase Tabla 13), sino que además fue seleccionada usando 11 criterios de filtro (depuración), de modo que ofrezca mejor validez a la fase de experimentación.

Tabla 1. Descripción de los predictores que conforman la matriz de diseño

No	Variables	Descripción
1	Mujer	1: Estudiante de género femenino (F); 0: Masculino
2	Soltero	1: Estudiante soltero; 0: Demás
3	Publica	1: Institución pública; 0: Privada
4	sem8.10	1: Entre 8 y 10 semestres cursados por el estudiante; 0: 11 o 12
5	bach.acad	1: Bachillerato académico; 0: Bachillerato técnico
6	hogar.habitual	1: Hogar actual permanente; 0: Temporal por estudio u otras
7	hogar.pers.max4	1: Hogar con máximo 4 personas; 0: Más de 4
8	cabeza.fam	1: Sí es cabeza de familia; 0: No
9	prof.pos.madre	1: Madre del estudiante profesional o con posgrado; 0: Grado inferior
10	prof.pos.padre	1: Padre del estudiante profesional o con posgrado; 0: Grado inferior
11	ocupa.madre.hogar	1: Madre del estudiante, principalmente dedicada a labores del hogar; 0: Principalmente se dedica a otras labores
12	estrat.estumin5	1: Vivienda donde vive el estudiante está en estrato 5 o 6; 0: Menor
13	sisbenfam.max2	1: Hogar del estudiante está clasificado en nivel de Sisbén de máximo 2; 0: Superior o no clasificado
14	pisos.max3	1: Material piso (habita el estud.) es tierra, arena o como mucho de madera burda, tabla/tablón; 0: Madera pulida, u otro superior.
15	internet	1: Hogar del estudiante cuenta con servicio de internet; 0: No
16	TV	1: Hogar del estudiante cuenta con servicio cerrado de televisión (cable, satelital, parabólica); 0: No
17	telef	1: Hogar del estudiante cuenta con servicio de teléfono fijo; 0: No
18	lavadora	1: Hogar del estudiante cuenta con lavadora habitual o permanente; 0: No
19	horno	1: Hogar del estudiante cuenta con horno eléctrico o a gas; 0: No
20	microon	1: Hogar del estudiante cuenta con horno microondas; 0: No
21	automov	1: Hogar del estudiante cuenta con automóvil particular; 0: No
22	ingres.fammax2	1: Ingresos mensuales del hogar del estudiante son de máximo 2 salarios mínimos legales; 0: Ingresos superiores
23	NOtrabaja.estu	1: El estudiante no labora para pagar la matrícula o ayudar con los gastos del hogar. 0: Sí la tiene.
24	edadmaxQ3	1: Edad del estudiante es inferior al cuartil 3 de las edades de la muestra; 0: es superior
25	g.Alto.comp.lecto	1: Alto desempeño del estudiante en comprensión lectora (puntaje mayor al cuartil 3). 0: Bajo desempeño (puntaje inferior al cuartil 1)
26	g.Alto.razo.cuanti	1: Alto desempeño del estudiante en Razonamiento cuantitativo (puntaje mayor al cuartil 3). 0: Bajo desempeño (puntaje inferior al cuartil 1)

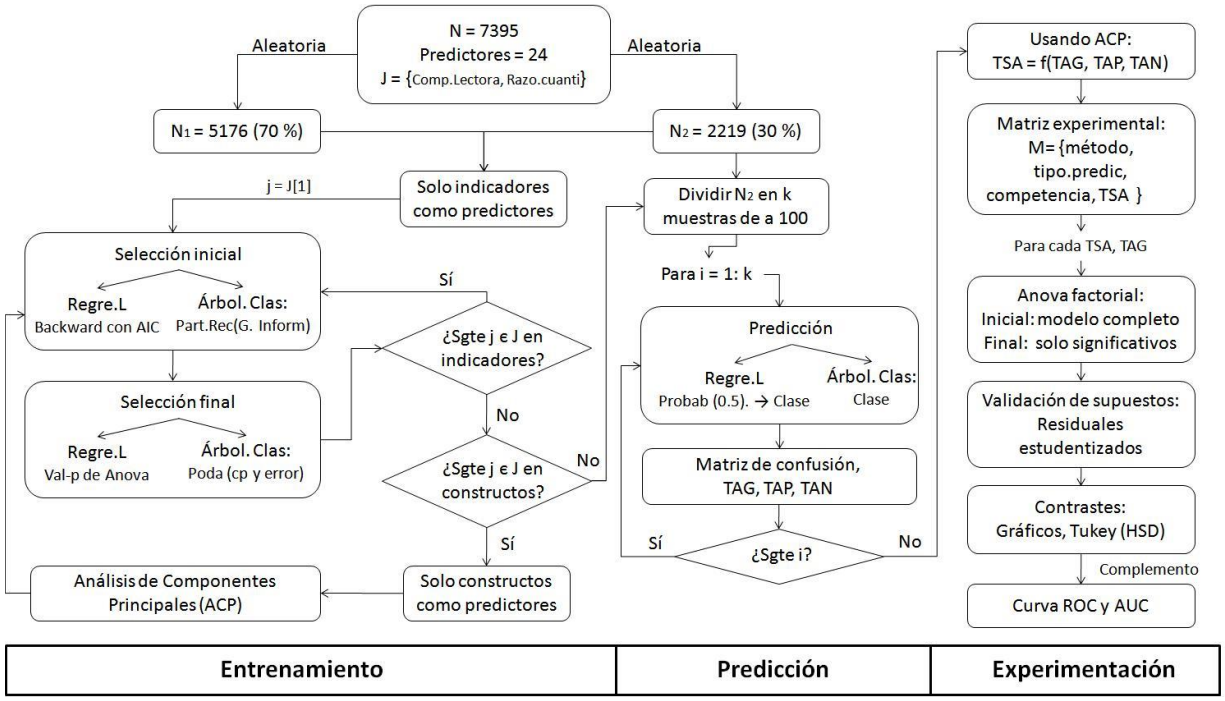


Fig. 1. Metodología EPE: Entrenamiento – Predicción – Experimentación

3.2. Depuración de la base de datos

En la tabla 1 se describe los 24 indicadores a considerar en el marco de los modelos a obtener y contrastar. Además, las dos últimas filas exponen las dos variables dependientes, por predecir.

3.3. Propuesta EPE: entrenamiento predicción experimentación

La figura 1 presenta la metodología propuesta. Nótese, en dicha figura, la articulación de conocimientos de aprendizaje estadístico, inteligencia artificial, diseño de experimentos y estadística multivariada.

La fase de entrenamiento constó del 70 % (5176) de la muestra y consistió en la generación de los modelos.

En la fase de predicción se probó los modelos sobre las instancias restantes (30 %; 2219), las cuales fueron divididas en 22 subgrupos (c/u con 100 observaciones). Así, por cada combinación método – competencia - tipo de predictor se obtuvo 22 observaciones de Tasa de Aciertos Global (TAG), Tasa de Aciertos Positivos (TAP) y Tasa de Aciertos Negativos (TAN), para un total de 176 observaciones.

La fase de experimentación partió de la construcción de la matriz experimental, en la cual se consolidaron dichas tasas y los factores de interés.

Además, desde una óptica multivariada, se combinaron TAG, TAN, TAP bajo ACP, para representar un nuevo índice: Tasa Subyacente de Aciertos (TSA).

La estrategia de experimentación se basó en un diseño factorial de tres factores. La tabla 2 resume la variable respuesta, los factores y niveles. Acorde con el diseño previo, el modelo de efectos del Análisis de Varianza se plasma en la ecuación 1:

$$TSA_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk} \ . \quad (1)$$

Tabla 2. Factores, respuesta y niveles

Factores	Niveles	Etiquetas
Método	2	Regre.L; Árbol
Tipo de predictores	2	Indicad; Construc
Competencia	2	Razo. Cuanti; Comp.Lecto
Tasa Aciertos Subyacente (TSA)	INF	0 – 1 (Respuesta)

Tabla 3. Resumen estadístico matriz de diseño

Variables	Mín	Máx	Prop
Mujer	0	1	0,32
Soltero	0	1	0,88
Publica	0	1	0,38
sem8.10	0	1	0,85
bach.acad	0	1	0,79
hogar.habitual	0	1	0,83
hogar.pers.max4	0	1	0,73
cabeza.fam	0	1	0,13
prof.pos.madre	0	1	0,28
profes.pos.padre	0	1	0,29
ocupa.madre.hogar	0	1	0,36
estrat.estumin5	0	1	0,11
sisbenfam.max2	0	1	0,22
pisos.max3	0	1	0,11
internet	0	1	0,92
TV	0	1	0,86
telef	0	1	0,87
lavadora	0	1	0,91
horno	0	1	0,63
microon	0	1	0,60
automov	0	1	0,44
ingres.fammax2	0	1	0,24
NOtrabaja.estu	0	1	0,42
edadmaxQ3	0	1	0,70
g.Alto.razo.cuanti	0	1	0,52
g.Alto.comp.lecto	0	1	0,51

Tabla 4. Modelos de regresión con indicadores para ambas competencias (70 %; 5176 instancias)

	g.Alto.comp.lecto	**g.Alto.razo.cuanti**
Mujer	-0,268***	-0,923***
	(0,066)	(0,070)
soltero	0,233**	
	(0,108)	
publica	1,235***	1,374***
	(0,069)	(0,073)
sem8.10	-0,450***	-0,734***
	(0,090)	(0,097)
prof.pos.madre	0,401***	0,475***
	(0,081)	(0,085)
profes.pos.padre	0,501***	0,551***
	(0,080)	(0,085)
estrat.estumin5	0,516***	0,791***
	(0,108)	(0,117)
sisbenfam.max2	-0,458***	-0,631***
	(0,078)	(0,082)
internet	0,342***	
	(0,115)	
ingres.fammax2	-0,283***	-0,373***
	(0,076)	(0,079)
NOtrabaja.estu		0,203***
		(0,069)
edadmaxQ3	0,623***	1,031***
	(0,076)	(0,076)
Constant	-1,049***	-0,478***
	(0,167)	(0,115)
Observ.	5.176	5.176
Log Likelihood	-3.155,015	-2.907,384

***Note:* *p<0,1; **p<0,05; ***p<0,01**

μ como la media global, común a todos los tratamientos; τ_i el efecto del i-ésimo método; β_j el efecto del j-ésimo tipo de predictor; γ_k el efecto de la k-ésima competencia; $(\tau\beta)_{ij}$ el efecto de la interacción método-tipo de predictor; $(\tau\gamma)_{ik}$ el efecto de la interacción método-competencia; $(\beta\gamma)_{jk}$ el efecto de la interacción tipo de predictor-competencia; ε_{ijk} el error aleatorio en las condiciones i, j, k; TSA_{ijk} como la respuesta obtenida bajo las condiciones i, j, k.

Dentro de la fase de experimentación, a modo de referencia de validez de criterio, se tomaron como criterios la Curva ROC y el Área Bajo la Curva (AUC, en inglés), los cuales son clásicos en la comparación de métodos de clasificación.Todos los análisis se ejecutaron en lenguaje R, por medio del entorno de RStudio.

Adicionalmente, se ha hecho uso de los siguientes paquetes:"MASS" (Venables, & Ripley, 2002), “car" (Fox, & Weisberg, 2011), "rpart" (Therneau, Atkinson & Ripley, 2017), “rpart.plot" (Milborrow, 2017), "gplots" (Warnes et al., 2016), "stargazer" (Hlavac, 2015) y "pROC" (Robin et al., 2011).

4. Resultados

4.1. Resumen estadístico de matriz de diseño

En la tabla 3 se resume los 24 indicadores y las dos variables dependientes, enfatizando sobre la proporción de ocurrencia de cada característica específica.

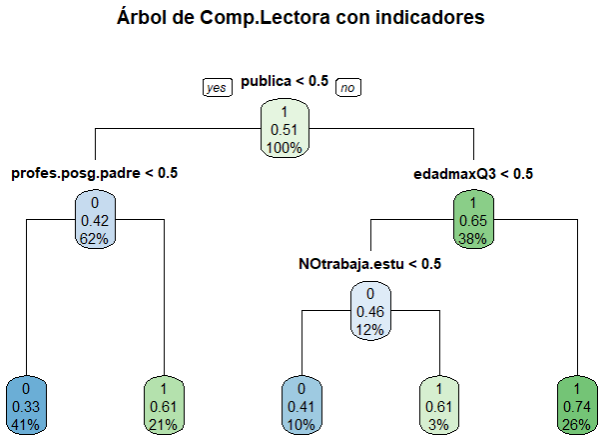


Fig. 2. Árbol de clasificación (solo con indicadores) para comprensión lectora

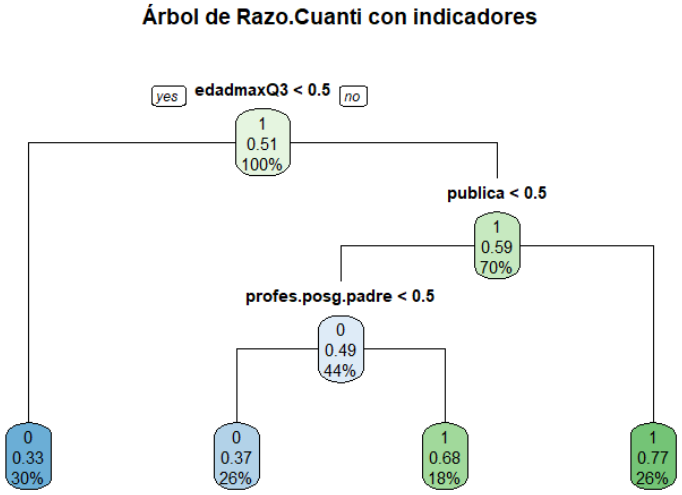


Fig. 3. Árbol de clasificación (solo con indicadores) para razonamiento cuantitativo

4.2. Regresión logística solo con indicadores

En la tabla 4, los modelos de regresión se presentan para ambas competencias. Se destaca que la mayoría de los predictores coinciden en ambos modelos, a excepción de soltero (solo en comprensión lectora resultó significativo), "Notrabajo.estu" (evento de que el estudiante no trabaje) y presencia de servicio de "internet" en el hogar. También, nótese que los signos de los coeficientes son idénticos en ambos modelos.

4.3. Árbol de clasificación solo con indicadores

Las figuras 2-3 ofrecen los árboles obtenidos por medio de la función *rpart*, siguiendo el método de partición recursiva con base en el índice de ganancia de información.

En la figura 2 puede verse que de los 24 indicadores en ensayo, solo 4 resultaron relevantes para discriminar las instancias respecto al alto/bajo desarrollo en comprensión lectora.

Tabla 5. Modelos de regresión con constructos

Comp.	g.Alto. comp.lecto	g.Alto.razo. cuanti
Comp.1	0,220*** (0,016)	0,302*** (0,017)
Comp.2	-0,179*** (0,021)	-0,246*** (0,022)
Comp.4	0,569*** (0,029)	0,794*** (0,033)
Comp.6	-0,067** (0,030)	
Comp.5		-0,085*** (0,031)
Comp.7	0,063** (0,030)	-0,065** (0,031)
Comp.8		0,146*** (0,032)
Comp.9	-0,063** (0,032)	-0,142*** (0,034)
Comp.11	0,161*** (0,033)	0,155*** (0,034)
Comp.14	-0,189*** (0,035)	-0,215*** (0,037)
Comp.15	-0,207*** (0,036)	-0,165*** (0,038)
Comp.18	0,076** (0,037)	0,087** (0,039)
Comp.19	-0,302*** (0,038)	-0,398*** (0,041)
Comp.20	-0,133*** (0,039)	-0,142*** (0,041)
Comp.22	0,090** (0,041)	0,208*** (0,043)
Constant	0,048 (0,030)	0,082** (0,032)
Observ.	5.176	5.176
Log. Likelihood	-3.157,891	-2.901,572
Note:	*p<0,1; **p<0,05; ***p<0,01	

Este árbol está compuesto por un nodo raíz (institución pública/privada) y tres nodos intermedios (profes.posg.padre, edadmaxQ3, NOtrabaja.estu). En ambos árboles (figuras 2-3), considerando el coeficiente de penalización (cp), a medida que se incluyen nodos, el error desciende.

Es decir, no hay razones para podar la estructura de los árboles actuales.

4.4. Regresión logística solo con constructos

La tabla 5 presenta los modelos de regresión que incorporan como predictores solamente constructos derivados de ACP. Estos son teóricamente independientes, lo que mitiga problemas colinealidad. De los 15 predictores aglomerados entre los dos modelos finales, solo tres no coinciden: componentes 5, 6 y 8.

4.5. Composición de los constructos

En la tabla 6 se plasma la estructura interna de los componentes 1 y 4. Una interpretación cualitativa a los constructos no es objeto de este estudio, pero, a modo de ejercicio, el componente 1 puede entenderse como un tipo de condición del estudiante para afrontar la formación universitaria, donde valores positivos representan facilitadores y, valores negativos, limitadores.

Véase, en la figura 4, la lógica del continuo bipolar "condiciones del estudiante". Considerando esta figura y la tabla 6, una posible interpretación para el coeficiente del componente 1 (0,302) en razonamiento cuantitativo es: aplicando el antilogaritmo a dicho coeficiente, se tiene que si la puntuación de las condiciones del estudiante (comp. 1) aumenta en una unidad, este es 1,35 veces más propenso a ubicarse en el grupo de alto desarrollo de competencias que en el bajo (asumiendo los demás constante).

Otra forma de entender los cambios en el alto/bajo desarrollo de competencias es mediante una tabla de frecuencias de dos vías (tabla 7). Entre los estudiantes con condiciones de vida limitadoras (puntajes menores de cero), el 42,2 % se ubica en alto desempeño (en razonamiento cuantitativo).

En cambio, entre los de condiciones facilitadoras (puntaje mayor de cero), esta proporción asciende al 60,3 %. Lo mismo en competencia lectora, donde se pasa de 44,3 % a 57,3 %.

En ambos casos, el test de independencia Chi-2 fue rechazado, lo cual favorece la creencia de una relación entre las variables (valores-p casi nulos). Nótese el valor interpretativo y la síntesis que propicia el enfoque de constructos en lugar de solamente indicadores observables.

Pasando al componente 4 (véase tabla 4), este podría interpretarse como otro tipo de condiciones del estudiante, que difieren de las del componente 1. En este caso, el estudiante cabeza de familia suma al indicador latente, así como: provenir de universidad pública, internet y que no trabaja. En cambio, género femenino, institución privada, más de 4 personas en el hogar, no es cabeza de familia, sí trabaja, entre otros, tiende a restar puntos al indicador latente.

Así, a través de los diferentes componentes y mediante un trabajo interdisciplinario que complemente el enfoque cuantitativo con interpretaciones desde las ciencias sociales y humanas, podría encontrarse diferentes perfiles de estudiantes. Para la asociación entre este constructo (4) y el alto/bajo desarrollo de competencias, se halló que, entre los estudiantes con condiciones limitadoras (puntajes negativos), el 41 % se ubicó en alto desarrollo en comprensión lectora; en cambio, para los de condiciones facilitadoras (puntajes positivos) este porcentaje aumentó a 64,6 %. Lo mismo sucedió en razonamiento cuantitativo, pues se pasó de 39,6 % a 67,7 %. De nuevo, tómese en cuenta la síntesis y el valor interpretativo del usar constructos, en vez de solo indicadores.

4.6. Árbol de clasificación solo con constructos

Las figuras 5-6 ofrecen los árboles para comprensión lectora y razonamiento cuantitativo.

Ningunos de estos árboles merece podarse, pues los errores tendieron a la baja a medida que aumentaron los nodos presentes.

A diferencia del árbol con constructos para competencia lectora, el árbol para razonamiento cuantitativo solo emplea los componentes 4 y 2 (excluye el 1). Vale decir que el componente 4 es suficiente para llegar a un nodo puro (hoja/terminal) de bajo desarrollo de competencias (0); estas representan el 37 % de la muestra de entrenamiento.

Tabla 6. Indicadores y cargas que definen los componentes 1 y 4

Indicadores	Comp.1	Comp.4
Mujer		-0,301
Soltero	0,162	
Publica	-0,162	0,564
sem8.10		-0,631
bach.acad	0,130	
hogar.habitual	0,105	-0,128
hogar.pers.max4		0,196
cabeza.fam	-0,188	0,113
prof.posg.madre	0,267	0,108
profes.posg.padre	0,282	0,102
ocupa.madre.hogar	-0,138	
estrato.estumin5	0,231	
sisben.fam.max2	-0,261	-0,166
pisos.max3	-0,155	-0,109
Internet	0,229	0,106
TV	0,242	
telef	0,187	
lavadora	0,224	
horno	0,283	
microon	0,256	
automov	0,312	

4.7 Comparativa experimental

En la fase experimental, primero fue necesario crear la TSA (Tasa Subyacente de Aciertos), la cual se forma por medio del ACP con indicadores TAG, TAN y TAP. En la tabla 8 se presentan las cargas de los indicadores de la matriz de confusión en cada componente.

Por motivos de análisis de los signos y considerando que una situación favorable deduce que los TAG, TAP y TAN incrementan y, a su vez, una situación desfavorable equivale a que estos disminuyan, se eligió el componente 2, en el que todas las cargas presentan el mismo signo (+).

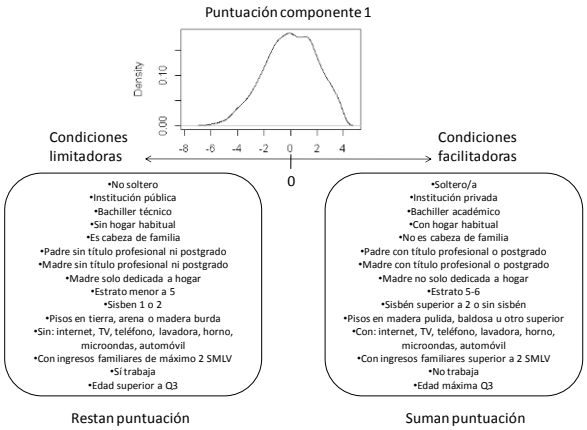


Fig. 4. Metodología EPE: Entrenamiento-Predicción – Experimentación

Tabla 7. Frecuencias de dos vías para categorías del componente 1

Comprensión Lectora		
Condiciones.comp.1	0: Bajo	1: Alto
0: Limitadoras	0,557	0,443
1: Facilitadoras	0,427	0,573
Test-Chi2, valor-p:	< 2,2e-16	
Razonamiento Cuantitativo		
Condiciones.comp.1	0: Bajo	1: Alto
0: Limitadoras	0,578	0,422
1: Facilitadoras	0,397	0,603
Test-Chi2, valor-p:	< 2.2e-16	

Tabla 8. Cargas de las componentes del ACP usando tasas de acierto de la matriz de confusión

Estad	**Comp.1**	**Comp.2**	**Comp.3**
TAG	0,113	0,558	0,822
TAN	-0,684	0,644	-0,343
TAP	0,721	0,524	-0,454
Des.Est	0,268	0,134	0,066
Propor	0,763	0,190	0,047

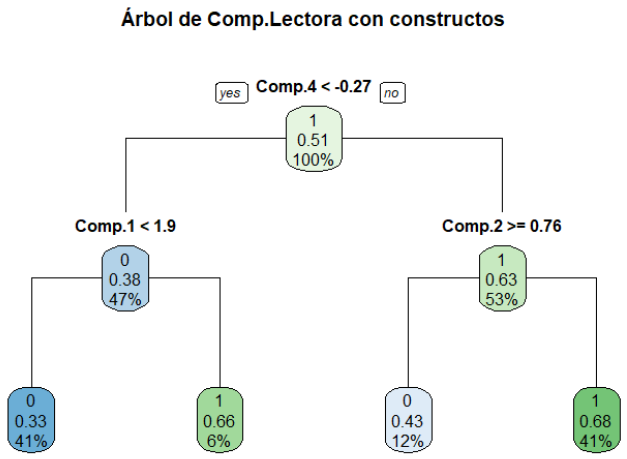


Fig. 5. Árbol de clasificación (solo constructos) para comprensión lectora

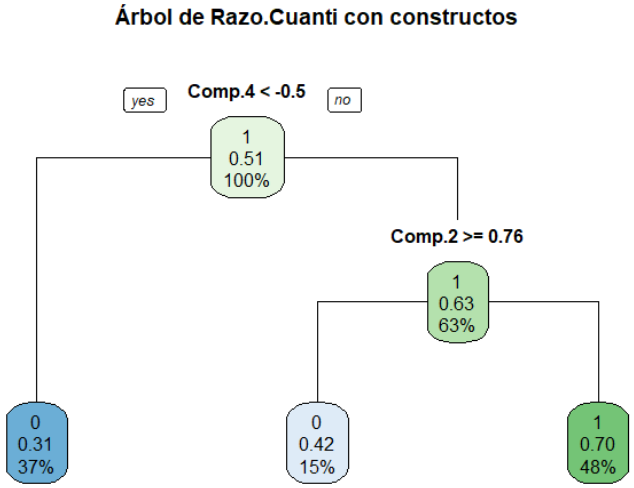


Fig. 6. Árbol de clasificación (solo constructos) para razonamiento cuantitativo

La correlación entre la TAG y la TSA es de 0,7668, que si bien es positiva y resultó significativa, la TSA considera también información de los otros dos indicadores de desempeño. Vale anotar que la TSA fue normalizada como proporción 0 - 1.

Análisis de varianza para TSA:

Luego de depurar el modelo Anova completo, según la significancia de los factores e interacciones, se llega al modelo expuesto en la tabla 9, que reporta como significativos (p-valores < 0,05) únicamente el método (Regresión Logística, Árbol de Clasificación), el tipo de predictor (indicadores, constructos) y la interacción. Este modelo resulta válido siempre y cuando se satisfagan los siguientes supuestos: error normal e independiente con media cero y varianza constante.

En la figura 7 se presentan diferentes gráficos de los residuales (estimaciones del error) que

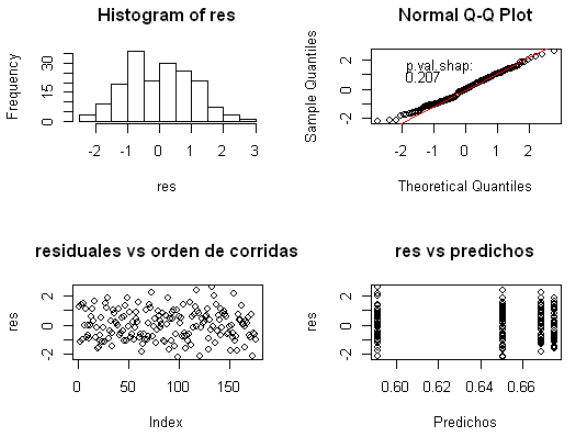


Fig. 7. Gráficos de residuales del Anova para TSA: normalidad, orden de corridas y predichos

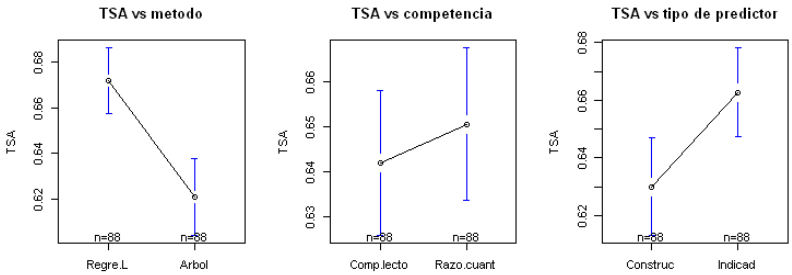


Fig. 8. Gráficos de medias de TSA según factores experimentales

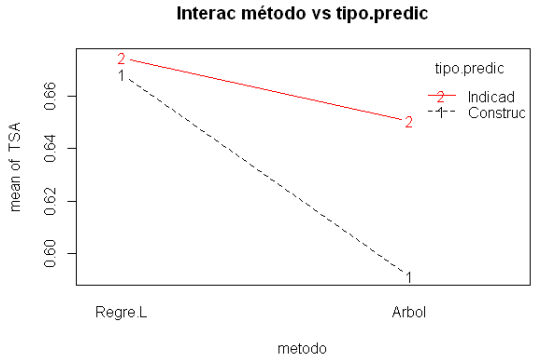


Fig. 9. Gráfico de interacción método-tipo de predictor

permiten explorar los supuestos de normalidad e independencia. En dicha figura se observa que no hay evidencia suficiente para rechazar normalidad e independencia de residuales. Además, bajo el test de Shapiro-Wilks, el valor p es muy superior a 0,1. Respecto a la homocedasticidad, tampoco se observa patrones en los residuales (estudentizados). Por consiguiente, los resultados obtenidos con Anova pueden considerarse razonables.

Tabla 9. Modelo final de Anova para TSA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Método	1	0,1137	0,11373	22,697	**4,01E-06**
tipo.predic	1	0,0473	0,04732	9,444	**0,00246**
Método:tipo.predic	1	0,031	0,03101	6,188	**0,01382**
Residuals	172	0,8619	0,00501		

Tabla 10. Comparaciones múltiples de Tukey al 95 % de confianza

Factores e interacciones	Niveles	Diff	p adj
Método	1. Arbol-Regre.L	-0,05084	**0,0000**
Tipo.predic	2. Indicad-Construc	0,0328	**0,0025**
Método-tipo.predic	3. Arbol:Construc-Regre.L:Construc	-0,0774	**0,0000**
	4. Regre.L:Indicad-Regre.L:Construc	0,0063	0,9760
	5. Arbol:Indicad-Regre.L:Construc	-0,0181	0,6304
	6. Regre.L:Indicad-Arbol:Construc	0,0836	**0,0000**
	7. Arbol:Indicad-Arbol:Construc	0,0593	**0,0007**
	8. Arbol:Indicad-Regre.L:Indicad	-0,0243	**0,3758**

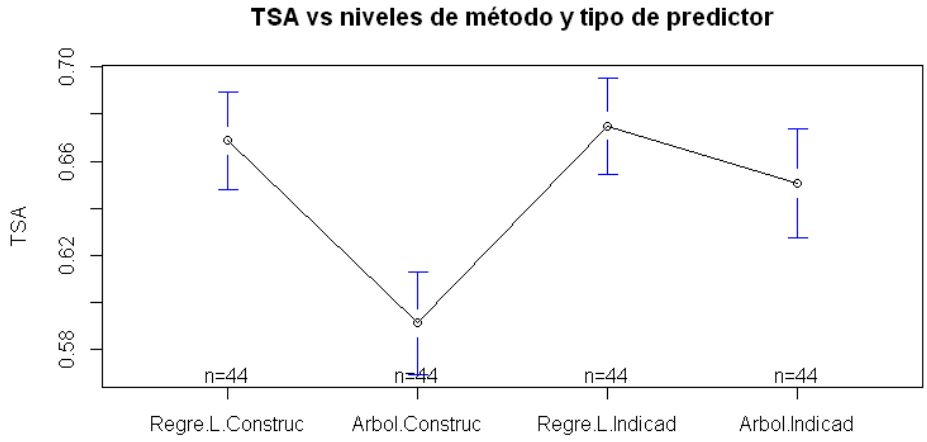


Fig. 10. Gráfico de medias de los niveles método-tipo_indicador para la TSA

Contraste de medias para TSA:

En la figura 8 se aporta las gráficas de medias según los factores individuales. Esta figura es un reflejo fiel de lo que dicta la tabla Anova para TSA. Para el factor competencia (comprensión lectora y razonamiento cuantitativo) los intervalos de confianza al 95 % para la media se solapan, lo que deduce que no hay evidencia suficiente para rechazar la hipótesis nula.

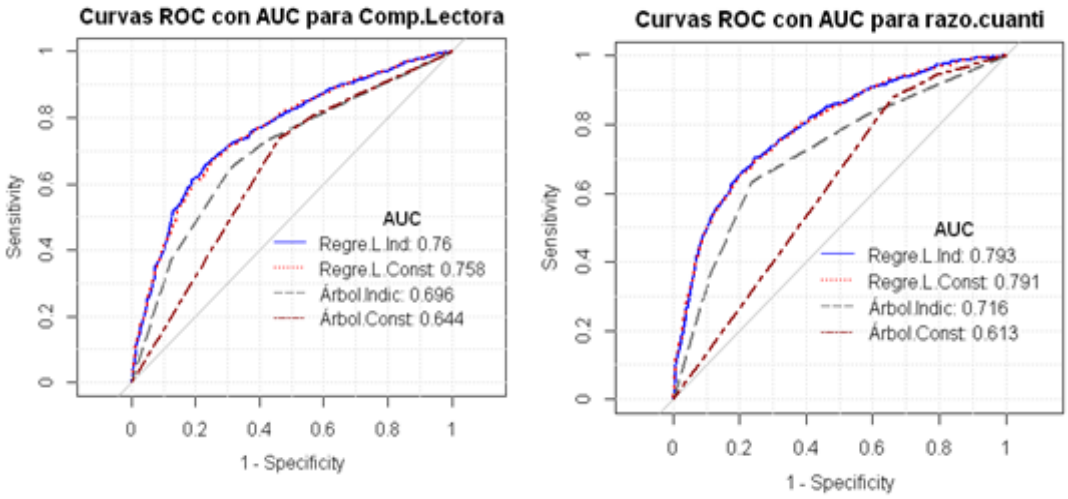


Fig. 11. ROC con AUC para comprensión lectora y razonamiento cuantitativo

Tabla 11. Test Delong para parejas de curvas ROC con AUC

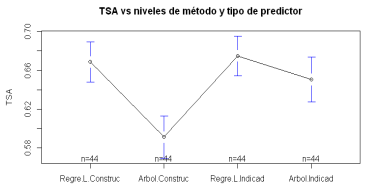
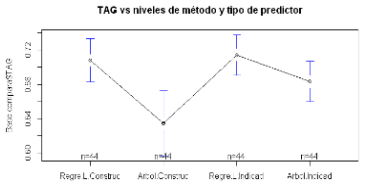
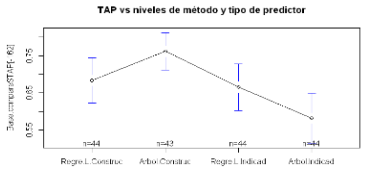
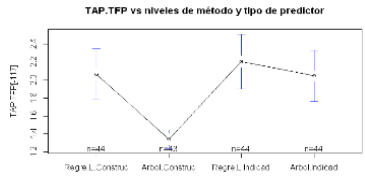
Comprensión lectora	Z	Valor-p	Conclusión
Regre.L Indi VS Regre.L Constr	10,053	0,3148	No hay diferencias
Árbol Indi VS Árbol Constr	51,167	3,11E-07	Diferencias. Árbol Indicad, mejor
Regre.L Ind VS Árbol Indi	8,968	2,20E-16	Diferencias. Regre.L, mejor
Regre.L Constr VS Árbol Constr	12,466	2,20E-16	Diferencias. Regre.L, mejor
Razonamiento cuantitativo	Z	Valor-p	Conclusión
Regre.L Indi. VS Regre.L Constr	11,011	0,2709	No hay diferencias
Árbol Indi VS Árbol Constr	92,557	2,20E-16	Diferencias. Árbol Indicad, mejor
Regre.L Indi VS Árbol Indicad	90,191	2,20E-16	Diferencias. Regre.L, mejor
Regre.L Constr VS Árbol Constr	19,04	2,20E-16	Diferencias. Regre.L, mejor

Contrario sucede para método y tipo de predictor, tal como lo reflejó la tabla 9. No obstante, recuérdese que se encontró presencia de una interacción entre método y tipo de predictor, y, de hecho, esta cobra más importancia que los efectos individuales, a la hora de concluir.

En la figura 9 se aporta el gráfico de interacción. Allí, se observa que la principal diferencia radica en el método árbol de clasificación cuando es usado con y sin constructos. Adicionalmente, en la tabla 10 se presenta los resultados del test de Tukey para comparar analíticamente la diferencia de medias entre pares de niveles de los factores. Entre Árboles de Clasificación y Regresión Logística la evidencia se inclina a favor de Regresión Logística respecto al desempeño (diferencia de -0,05084 y un valor p de 0,0000).

No obstante, las conclusiones sobre los efectos individuales pasan a nivel secundario cuando hay interacción, como ocurre con la combinación método - tipo de predictor. Dichas diferencias ocurren en (véase tabla 10): Árboles de Clasificación con constructos contra Regresión con constructos (valor-p: 0,0000); Regresión Logística con indicadores contra Árboles de Clasificación con constructos (valor-p: 0,0000); y Árboles con indicadores contra.

Tabla 12. Resumen de las principales conclusiones y hallazgos de Anova para TSA y diversas variables de la matriz de confusión

Var	Normal	Homocedast.	Transf.	Normal	Homoced	Conclusión	Gráficos
TSA	Sí (0,207)	Sí (Método: 0,59; Tipo.predic: 0,57; Competencia: 0,26	No	NA	NA	Interacción método*tipo.predic. Árbol con Constructos es inferior a los demás	
TAG	Sí (0,297)	No (Método: 0,15; Tipo,predic: **0,005**; Competencia: 0,78)	Arcsin (y)	Sí (0,106)	No (Método: 0,59; Tipo,predic: **0,021**; Competencia: 0,78)	Método y tipo.predic. Reg. superior a Árbol. Indicad superior a Constructos. Árbol.Indicad vs Árbol.Const valor-p de 0,0504.	
TAP	No (0,0001)	No (Método: 0.325; Tipo.predic: 0.027; Competencia: 0.36)	arcsin(y)	No (0.006)	Sí (Método: 0,61; Tipo.predic: 0,72; Competencia: 0,54)	Método* tipo.predic. Árbol con Constructos es superior a Árbol sin constructos	
TAN	No (0.0000)	Sí (Método: 0,23; Tipo,predic: 0,84; Competencia: 0,36	arcsin(y)	No (0,004)	Sí (Método: 0,09; Tipo.predic: 0,21; Competencia: 0,35)	Método*tipo.predic y tipo.predic*competencia . Árbol con Constructos es inferior a los demás	
TAP / TFP	No (0.0000)	No (Método: 0,028; Tipo,predic: 0,001; Competencia: 0,59	log(y)	No (0,038)	No (Método: 0,20; Tipo.predic: 0,0003; Competencia: 0,47)	Método*tipo.predic. Árbol con Constructos es inferior a los demás.	

Árboles de Clasificación con constructos (valor-p: 0.0007).

Las demás no presentan diferencias estadísticamente significativas. En la figura 10 se percibe mejor las diferencias en TSA. Para Regresión Logística (con indicadores o constructos), los intervalos de confianza para la diferencia de medias se solapan, lo mismo que entre estos y Árboles de Clasificación con indicadores. Sin embargo, Árbol de Clasificación con constructos difiere notablemente de las demás combinaciones.

Es decir, el uso de constructos en Árboles de Clasificación, en este contexto específico, deterioró considerablemente la TSA, lo que no ocurrió en Regresión Logística.

Tabla 13. Resumen de comparativas Regresión Logística Vs. Árboles de Clasificación

Estudio	Contexto	Variable respuesta	Tamaño de muestra	Otros métodos	Ganador	Observaciones
Fernandez et al. (2016)	Salud	Madres con alto riesgo a contraer mastitis	368	NO	RL - AC	Regresión logística: mejor desempeño en ROC, precisión y sensibilidad. Árboles de regresión: mejor especificidad
Coussement et al. (2014)	Marketing	Evento de que los clientes respondan o no a un correo de marketing	195.751	SI	AC	Buscaron comparar el desempeño con y sin datos imprecisos. Dividen el conjunto de datos en cinco grupos de igual tamaño.
Grochowska et al. (2014)	Zoología	Tamaño de las camadas paridas por especie “oveja polaca”	323	NO	AC	
Demir (2014)	Salud	Reincidencia o no del paciente, dentro de los 45 días desde que se le dio de alta	963	SI	RL	Ambos presentaron desempeño superior. Diferencias despreciables. Aunque fue levemente más alto el de Regresión Logística
Dieguez et al. (2014)	Finanzas	Evento de que una pequeña empresa (o micro) sea liquidada por insolvencia económica	39710	NO	AC	
Chen et al. (2014)	Química	Biodegradación o no de compuestos químicos	1.629	NO	AC	
Yahya and Ismail (2014)	Botánica	Mortalidad de los árboles en los bosques de lluvia tropical en Camboya	383	SI	AC	
Safiarian et al. (2013)	Salud	Alto riesgo de sangrado luego de cirugía de bypass de arteria coronaria	205	NO	AC	

Comparación de modelos bajo curva ROC y AUC:

La curva ROC con AUC es el criterio más común para comparar clasificadores. En la medida en que un método más se aleje de la diagonal (AUC: 0.5), hacia arriba e izquierda, se considera de mejor desempeño. En la figura 11 se muestra las curvas ROC con AUC para ambas competencias. Para explorar la significancia de las diferencias percibidas en la figura 11, la tabla 11 presenta el test DeLong para parejas de curvas ROC. La Regresión Logística muestra mejor desempeño que Árboles de Clasificación, tanto con indicadores como con constructos.

Entre Regresión logística con indicadores y con constructos no existe diferencia significativa. En cambio, Árboles de Clasificación muestra mejor desempeño con indicadores que con constructos, lo cual es consistente con lo obtenido bajo Anova para TSA. Cuando se compara Regresión Logística (indicadores o constructos) contra Árboles de Clasificación con indicadores, los resultados de ambos criterios difieren. Bajo Anova para TSA no se encuentran diferencias significativas, en cambio, bajo Curvas ROC con AUC se mostró superior Regresión Logística. Estas discrepancias motivan explorar otras métricas de la matriz de confusión.

La tabla 12 expone el resumen de lo obtenido con Anova para TSA, pero además para TAG, TAP, TAN y la razón entre TAP/TFP. Estas nuevas comparaciones de factores e interacciones bajo Anova es una labor meramente exploratoria, a excepción de lo expuesto para TSA, pues en la mayoría de los casos no se cumplió al menos uno de los supuestos de los residuales. De hecho, a diferencia de TSA, para las demás variables fue necesario transformarlas, a fin de mitigar en cierta medida la violación de supuestos. Lo importante con este ejercicio es notar hacia dónde apuntan los nuevos hallazgos.

Entonces, con base en la tabla 12 y las gráficas de medias, los resultados para TAG, TAN y TAP/TFP son consistentes con lo que reflejó Anova para TSA: Regresión Logística (indicadores o constructos) y Árboles de Clasificación (con indicadores) presentan un desempeño semejante, que además es superior al de Árboles de Clasificación con constructos.

En cambio, en la variable TAP el uso de constructos en los Árboles de Clasificación mejoró sustancialmente el desempeño, siendo superior al de Árbol con indicadores. Todo esto refleja lo complejo y subjetivo que puede ser concluir sobre el desempeño de clasificadores con base en un solo criterio.

5. Conclusiones

Para la primera pregunta de investigación, de si el uso exclusivo de constructos en los métodos de clasificación mejora el desempeño de estos, se encontró lo siguiente: para la TSA bajo Regresión Logística, ni Anova ni Curvas ROC con AUC reflejaron diferencias en el desempeño de los métodos (con constructos o indicadores). Sin embargo, en Árboles de Clasificación se notó un detrimento de la TSA y de la curva ROC con AUC cuando se usó constructos, en vez de indicadores. Respecto a la TAP, el uso de constructos tendió a mejorar el desempeño de Árboles de Clasificación, en contraste con el uso de indicadores.

En general, vale resaltar la capacidad de interpretación al usar constructos, pues hizo posible encontrar patrones de condiciones del estudiante, que pueden verse como facilitadores/limitadores para su desarrollo de competencias; algo para aprovechar en futuros estudios, sobre todo en ciencias sociales y humanas.

Respecto a la segunda pregunta, sobre qué método presenta mejor desempeño, en el escenario de solo indicadores la evidencia de Anova TSA fue contundente hacia un empate entre Regresión Logística y Árboles de Clasificación. Esto también es visto en hallazgos (exploratorios) de Anova para TAP, TAN, TAG y TAP/TFP, lo cual difiere de los hallazgos en curva ROC con AUC, que mostraron ganador a Regresión Logística. En el escenario de solo constructos, tanto la evidencia de Anova para TSA como de Curvas ROC con AUC pusieron muy por encima a la Regresión Logística.

Para el tercer interrogante, sobre los principales predictores, para solo indicadores prevalecieron como significativos en ambos modelos de Regresión Logística (Comprensión Lectora y Razonamiento Cuantitativo):

Tabla 14. Documentación del paso a paso para la selección de la muestra de instancias

No	Filtro (F) / Adición (A)	Criterio	Especificaciones	No. Obs.
1	A	Cargar archivo 20143	SBPRO-20143-RGSTRO-CLFCCN-U_GEN.csv	209.726
2	A	Cargar archivo 20153	SBPRO-20153-RGSTRO-CLFCCN-U_GEN.csv	324.849
3	A	Conformar archivo completo	Archivos 20153 y 20143; se adiciona *period.SPRO*	**534.575**
4	F	Municipio de presentación de examen	Ciudades principales: Cali, Bogotá, Medellín, Barranquilla	265.453 (49,7%)
5	F	Municipio de residencia	Ciudades principales: Cali, Bogotá, Medellín, Barranquilla	146.284 (27,4%)
6	F	Estudiante local	Que resida en el mismo municipio donde presenta el examen; variable de control *reside.presenta*	144.647 (27,1%)
7	F	Programas académicos	Programas de ingeniería	26.786 (5%)
8	F	Naturaleza de la institución	Universidad o institución universitaria; variable *univers*	25.032 (4,9%)
9	F	Sector de institución	Solo oficial o no oficial (Ej: se excluye régimen especial); variable pub.priv	25.032 (4,9%)
10	F	No reporte discapacidad	Sin reporte de alguna discapacidad; variable control *alguna.disc*	17.487 (3,3%)
11	F	Sin puntajes de cero	Sin puntajes cero (Comprensión lectora y Razonamiento cuantitativo); variables: y.comp.lectora, y.razo.cuanti	17.479 (3,3%)
12	F	Número de semestres cursados	Entre 8 y 12 semestres	16.876 (3,2%)
13	A	Horas trabajadas por el estudiante	Variables *horas.trab.est* (incluye el cero en caso de que no trabaje)	16.876 (3,2%)
14	A	Edad	Variable edad en años hasta el 10/Jul/2017 (entera)	16.876 (3,2%)
15	A	g.Alto.comp.lecto; g.Alto.razo.cuanti	1: Estudiante con “Alto desempeño” en determinada competencia (puntaje superior al cuartil 3). 0: “Bajo desempeño” (puntaje inferior al cuartil 1)	16.876 (3,2%)
16	F	Alto / Bajo desempeño	Se excluyen los estudiantes con puntajes mayores a cuartil 1 y menores a cuartil 3	7.426 (1,39%)
17	F	Casos completos tomando en cuenta la matriz de diseño (véase tabla 1)	24 posibles predictores binarios, donde 1 es la presencia de la característica y 0 la ausencia de esta	**7.395** (1,38%)

Género (mujer: coeficientes negativos; hombre: +), tipo de institución (pública: +; privada: -), número de semestres al momento de la prueba (8-10: +; 11-12: -), grado profesional universitario o posgrado de madre (lo mismo para el padre) (sí: +; no: -), estrato del hogar del estudiante (5-6: +; 1-4: -), nivel de sisbén (1-2: -; 3 o más o ninguno: +) ingresos en el hogar del estudiante (1-2 SMLV: -; más de 2 SMLV: +), edad máxima del estudiante (inferior al cuartil 3 de la muestra: +; superior: -). Sin considerar el género, los niveles con signo positivo reflejan condiciones de vida tal vez más facilitadoras que limitadoras para la vida académica. Para Árboles de Clasificación, los que mejor ayudaron a discriminar y que coinciden en ambas competencias fueron tres, ya expuestos en

Regresión Logística: tipo de institución, grado profesional universitario/posgrado de padre, y edad máxima Q3. Sobre constructos, aquellos que coinciden en ambas competencias y métodos son los componentes 1, 2 y 4. Estos aportan más información cualitativa que los indicadores por separado, pues reflejan condiciones o perfiles del estudiante que tienen relación con el alto/bajo desarrollo de competencias. Es decir, aportan información para interpretaciones más rigurosas desde las ciencias sociales y humanas, a las que no es posible llegar solo con indicadores.

Para la cuarta pregunta, de si los predictores difieren según la competencia, los resultados se muestran consistentes (con indicadores o constructos), independiente de cuál es la competencia. Es de señalar que la Regresión incorporó más predictores significativos que los Árboles de Clasificación.

Este trabajo también aporta un procedimiento de comparación de métodos, que puede ser complementario al clásico de Curvas ROC con AUC, a fin de ayudar al analista, sobre todo ante conclusiones en conflicto.

Este estudio, además, enfatiza en la documentación del procedimiento de preparación y limpieza de datos, como un requisito crítico, de alto consumo de recursos, antes de llevar a cabo los análisis. Este comúnmente es obviado, pero desde el paradigma de investigación reproducible está tomando cada vez más fuerza. Otro aporte adicional del estudio es que provee una nueva matriz de datos experimentales, originales, que sirve para uso docencia e investigación, de modo que se nutran los procesos de formación con datos reales y debidamente controlados, así como que dé lugar a la creación de nuevas variables o enfoques comparativos.

Como trabajos futuros, vale tomar en consideración que este estudio solo abordó dos tipos de competencias en ingeniería. Nuevos trabajos deberían ampliar el alcance a otras competencias de Saber Pro, así como incorporar otras carreras, no solo de ingeniería, para validar la posibilidad de generalización. Asimismo, independiente del contexto, convendría que combinar y comparar las conclusiones obtenidas bajo Curvas ROC con AUC contra el enfoque propuesto de Anova con TSA (con uso de Análisis de Componentes Principales).

Otra oportunidad es aprovechar la riqueza interpretativa que posibilita el uso de constructos, al sintetizar cantidad de información proveniente de indicadores observables en las pruebas de Estado. Otra oportunidad es crear nuevas métricas que sirvan como variables respuesta para comparar con los resultados de la TSA y efectuar otros procedimientos para afrontar el eventual incumplimiento de supuestos.

Estudio emplea métodos tradicionales de estadística, como el diseño experimental y el Anova, desde un enfoque de Ciencia de Datos que los articula con conocimientos de Ciencias de la Computación, para automatizar tareas desde la preparación hasta la visualización de datos. Ello resulta fundamental en la actualidad, tomando en cuenta el fenómeno del Big Data, que hace inviable el procesamiento manual o bajo software tradicional. Futuros estudios en Ciencia de Datos y afines, convendría que aprovecharan el potencial del diseño experimental que provee la estadística clásica. Como se ha visto en este trabajo, así sea con datos observacionales y con abundante cantidad de estos, es posible generar nuevas matrices experimentales y complementar las comparativas entre clasificadores u otros métodos.

Agradecimientos

Al Instituto Colombiano para la Evaluación de la Educación ICFES (proveedor de los datos). A docentes y directivos del Máster en Análisis y Visualización de Datos Masivos de UNIR. Al docente Juan Carlos Correa (Ph.D) de la Escuela de Estadística de la Universidad Nacional de Colombia, Medellín.

Referencias

1. **Macina, O. (2004).** The utility of structure–activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models. *SAR and QSAR, Environmental Research,* Vol. 15, No. 1, pp. 1–18. DOI: 10.1080/1062936032000169633.
2. **Bahamón, M. & Reyes, L. (2014).** Caracterización de la capacidad intelectual, factores sociodemográficos y académicos de estudiantes

con alto y bajo desempeño en los exámenes Saber Pro-año 2012. *Avances en Psicología Latinoamericana,* Vol. 32, No. 3, pp. 459–476. DOI: 10.12804/apl32.03.2014.01.

3. **Caicedo, E., Guerrero, S., & López, D. (2016).** Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro*. Comunicaciones en Estadística,* Vol. 9, No. 1, pp. 93–106. DOI: 10.15332/s2027-3355.2016.0001.05.
4. **Cañon, M. & Jimenez, S. (2017).** Enfrentando resultados programa de ingeniería de sistemas de la USB con las pruebas Saber Pro. *Revista Investigación y Desarrollo en TIC,* Vol. 3, No. 1.
5. **Osma-Castellanos, W. A., Mojica-Perdomo, A. D., & Rivera-Florez, T. E. (2014).** Factores asociados al rendimiento en las Pruebas Saber Pro en estudiantes de Ingeniería Civil en universidades colombianas. *Revista Innovaciencia,* Vol. 2, No. 1. DOI: 10.15649/2346075X.234.
6. **Chen, G., Li, X., Chen, J., Zhang, Y. N., & Peijnenburg, W. J. (2014).** Comparative study of biodegradability prediction of chemicals using decision trees, functional trees, and logistic regression. *Environmental toxicology and chemistry,* Vol. 33, No. 12, pp. 2688–2693. DOI: 10.1002/ETC.2746.
7. **Chou, H. L., Yao, C. T., Su, S. L., Lee, C. Y., Hu, K. Y., Terng, H. J., & Wahlqvist, M. L. (2013).** Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees. *BMC bioinformatics,* Vol. 14, No. 1, pp. 100. DOI: 10.1186/1471-2105-14-100.
8. **Cortés, S. & Piñeros, C. (2015).** Examen de calidad de la educación superior Saber Pro: factores sociodemográficos en el desempeño académico en Instrumentación Quirúrgica. *Repert. med. cir,* Vol. 24, No. 3, pp. 206–211.
9. **Coussement, K., Van den Bossche, F., & De Bock, K. (2014).** Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research,* Vol. 67, No. 1, pp. 2751–2758. DOI: 10.1016/j.jbusres.2012.09.024.
10. **Demir, E. (2014).** A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. *Decision Sciences,* Vol. 45, No. 5, pp. 849–880. DOI:10.1111/deci.12094/full.
11. **Dieguez, A. I., Blanco-Oliver, A., & Vazquez-Cueto, M. J. (2015).** A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models. *Procedia Economics and Finance,* Vol. 23, pp. 9–14. DOI: 10.1016/S2212-5671(15)00797-2.
12. **Felicísimo, Á., Cuartero, A., Remondo, J., & Quirós, E. (2013).** Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides,* Vol. 10, No. 2, pp. 175–189. DOI:10.1007/s10346-012-0320-1.
13. **Fernández, L., Mediano, P., García, R., Rodríguez, J., & Marín, M. (2016).** Risk Factors Predicting Infectious Lactational Mastitis: Decision Tree Approach versus Logistic Regression Analysis. *Maternal and child health journal,* Vol. 20, No. 9, pp. 1895–1903. DOI: 10.1007/s10995-016-2000-6.
14. **Fernández, Y. (2011).** Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. *Investigación educativa,* Vol. 15, No. 27, pp. 165–180.
15. **Fox, J. & Weisberg, S. (2011).** *An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.*
16. **Gil, F., Rodríguez, V. A., Sepúlveda, L. A., Rondón, M. A., & Gómez-Restrepo, C. (2013).** Impacto de las facultades de medicina y de los estudiantes sobre los resultados en la prueba nacional de calidad de la educación superior (SABER PRO). *Revista Colombiana de Anestesiología,* Vol. 41, No. 3, pp. 196–204. DOI: 10.1016/j.rca.2013.04.003.
17. **Grochowska, E., Piwczyński, D., Portolano, B., & Mroczkowski, S. (2014).** Analysis of the influence of the PrP genotype on the litter size in Polish sheep using classification trees and logistic regression. *Livestock Science,* Vol. 159, pp. 11–17. DOI: 10.1016/j.livsci.2013.11.008.
18. **Hlavac, M. (2015).** *stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2. http://CRAN.R-project.org/package=stargazer*
19. **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015).** *An Introduction to Statistical Learning: With Applications in R.* Springer.
20. **Khemphila, A. & Boonjing, V. (2010).** Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. *International Conference on Computer Information Systems and Industrial Management Applications (CISIM'10),* pp. 193–198. DOI: 10.1109/CISIM.2010.5643666.

21. **Kurt, I., Ture, M., & Kurum, A. T. (2008).** Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications,* Vol. 34, No. 1, pp. 366–374. DOI: 10.1016/j.eswa.2006.09.004.
22. **León, A., Amaya, S., & Orozco, D. (2012).** Relación entre comprensión lectora, inteligencia y desempeño en pruebas Saber Pro en una muestra de estudiantes universitarios. *Cultura, Educación y Sociedad,* Vol. 3, No. 1.
23. **Milborrow, S. (2017).** *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 2.1.2. https://CRAN.R-project.org/ package =rpart.plot.*
24. **Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011).** Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications,* Vol. 38, No. 12, pp. 15273–15285. DOI: 10.1016/j.eswa.2011.06.028.
25. **Ramírez, C. (2014).** Factores asociados al desempeño académico según nivel de formación pregrado y género de los estudiantes de educación superior Colombia. *Revista Colombiana de Educación,* Vol. 66, pp. 203–224.
26. **Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011).** pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, Vol. 12, No. 77. DOI: 10.1186/1471-2105-12-77.
27. **Romero, D., Villarreal, E., & Velandia, N. (2015).** Resultados en Saber Pro de estudiantes de modalidad presencial y virtual en dos universidades colombianas. *Revista Academia y Virtualidad,* Vol. 8, No. 2, pp. 100. DOI: 10.18359/ravi.1426.
28. **Safiarian, R., Amini, P., Moez, E., Mohammadzadeh, F., Tavakoli, M., & Zayeri, F. (2013).** Risk group classification for bleeding after coronary artery bypass graft surgery: a comparison of the logistic regression with decision tree models. *Türk Göğüs Kalp Damar Cerrahisi Dergisi,* Vol. 21, No. 3, pp. 574–580.
29. **Sugimoto, M., Takada, M., & Toi, M. (2013).** Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer. *IEEE Conference of the Engineering in Medicine and Biology Society (EMBC'13), 35th Annual International,* pp. 3054–3057. DOI: 10.1109/EMBC.2013.6610185.
30. **Therneau, T., Atkinson, B., & Ripley, B. (2017).** rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart, Vol. 4, pp. 1–11.
31. **Torres, M., Vélez, J., & Altamar, F. (2015).** La calidad de la educación superior en Colombia. Una aproximación econométrica. *Clío América,* Vol. 9, No. 18, pp. 143–156.
32. **Vargas, G. (2007).** Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista educación,* Vol. 31, No. 1, pp. 43–63.
33. **Venables, W. & Ripley, B. (2002).** Modern Applied Statistics with S. Fourth Edition.
34. **Warnes, G., Bolker, B., Bonebakker, L., Gentleman, R., Andy, W., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2016).** *Various R Programming Tools for Plotting Data.* R package version 3.0.1.
35. **Yahya, Y., Ismail, R., Vanna, S., & Saret, K. (2014).** Using data mining techniques for predicting individual tree mortality in tropical rain forest: logistic regression and decision trees approach. *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, Vol. 91. DOI: 10.1145/2557977. 2557989.
36. **Zayeri, F., Seyedagha, S., Aghamolaie, H., Boroumand, F., & Yavari, P. (2016).** Comparison of the Logistic Regression and Classification Tree Models in Determining the Risk Factors and Prediction of Breast Cancer. *Iranian Journal of Epidemiology,* Vol. 12, No. 2, pp. 49–57.

Article received on 16/09/2017; accepted on 28/02/2018. Corresponding author is Jorge Pérez-Rave.