Encuesta Internacional sobre el abandono en la Educación Superior

AUTOR: Proyecto ALFA GUIA DCI-ALA/2010/94
FECHA: 29 de Abril de 2014

Este proyecto ha sido financiado con la ayuda de la unión Europea. Los contenidos son responsabilidad de sus autores, y de ningún modo debe considerarse que refleja la opinión oficial de la Unión Europea.

Proyecto ALFA-III

“Gestión Universitaria Integral del Abandono”

Aplicación de la teoría de clasificación al problema del abandono estudiantil: caso Universidad de Antioquia

Autores: Santiago Gallón (Universidad de Antioquia)
Johanna Vásquez (Universidad de Antioquia)
Grupo Análisis

Proyecto ALFA GUIA DCI-ALA/2010/94

(29 de Abril de 2014)

Este documento se ha realizado con la ayuda financiera de la Unión Europea.
El contenido de este documento es responsabilidad exclusiva de sus autores y en modo alguno debe considerarse que refleja la posición de la Unión Europea.

Aplicación de la teoría de clasificación al problema del abandono estudiantil: caso Universidad de Antioquia

Santiago Gallón*[a,c] and Johanna Vásquez†[b]

[a] *Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia.*
[b] *Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia.*
[c] *Institut de Mathématiques de Toulouse, Université Toulouse III Paul Sabatier, Toulouse, France.*

29 de abril de 2014

1. Introducción

Uno de los problemas a los que se enfrentan constantemente las directivas de una institución educativa, sea ésta de nivel primario, secundario, técnico-tecnológico o universitario, de carácter público o privado, es el relacionado con el fenómeno de la permanencia de sus estudiantes en los cupos ofrecidos. En particular, para el caso de las Instituciones de Educación Superior −IES−, éstas han venido presentando gran preocupación debido a que, ante el aumento en la demanda y la política ampliación de cobertura de la educación superior que ha caracterizado los recientes años, el número de estudiantes que logran culminar sus estudios es mínimo, evidenciándose un abandono de un gran número de estudiantes, principalmente, durante los primeros semestres. Las causas y consecuencias socio-económicas del abandono académico han sido ampliamente estudiadas en la literatura académica nacional e internacional y, aunque muchas aproximaciones conceptuales han surgido para explicar porqué un estudiante toma la decisión de abandonar sus estudios a lo largo de su ciclo académico (véase, por ejemplo, Spady [20], Tinto [22], Bean [1], Cabrera et al. [3] y Castaño et al. [9], y los trabajos de Castaño et al. [8, 7, 6] realizados para la Universidad de Antioquia, y las referencias allí citadas), no se dispone de muchas aproximaciones empíricas que permitan su correcta clasificación y predicción.

En este sentido, la investigación relacionada con el abandono estudiantil podría dividirse en dos grandes corrientes: *a*) en estudios que apuntaban a la profundización teórica del problema a partir del análisis metodológico de sus determinantes, es decir del efecto a sus causas, y *b*) en aquellos estudios interesados en encontrar la manera

*santiagogallon@gmail.com
†jovasve@gmail.com

más robusta de medir sus determinantes y niveles de riesgo, es decir profundización metodológica. De la primera, se destaca el avance en la definición del abandono con respecto a sus dimensiones temporal y espacial, la diferenciación entre abandono permanente y transitorio, las fases y niveles de riesgo de abandono, el impacto de los primeros cursos matriculados, y la interacción dinámica de sus constructos, entre otros. De la segunda corriente, se nota el paso de la evaluación económica a la aplicación de modelos de regresión con variable dependiente discreta tales como los modelos logit y probit, modelos multinomiales, la aplicación de modelos de duración (también de supervivencia) y la tendencia hacia regresiones multinivel. En estos estudios se analiza el abandono como resultado de la interacción de diferentes determinantes (véase por ejemplo, Pascarella and Terenzini [17], Willett and Singer [23], Booth and Satchell [2], Cameron and Heckman [4], Porto and Di Gresia [18], Cameron and Taber [5], DesJardins et al. [11], Cornwell [10], DesJardins et al. [12], Giovagnoli [14], Häkkinen and Uusitalo [15] y Ruthaychonnee [19]).

En consecuencia, dado que el abandono estudiantil es considerado como uno de los factores que más incide en el acceso y cobertura de la educación, su medición, estudio y monitoreo debe ser parte de los continuos procesos de evaluación de la eficiencia del sistema educativo, de ahí que es imperativo que las instituciones diseñen políticas y/o programas para disminuir o controlar el abandono estudiantil. En este sentido, la encuesta sobre causas y decisiones de abandono de estudios de educación superior del Proyecto Alfa-GUÍA (Gestión Universitaria Integral del Abandono, http://www.alfaguia.org/) ha intentado recoger información a nivel internacional que permita conocer mejor las causas que motivan el abandono con el fin de contribuir al diseño de estrategias que contribuyan a su reducción.

A partir de la forma como se diseñó la encuesta y de la información recolectada a partir de la aplicación de la misma por parte del Proyecto Alfa-GUÍA, ha sido posible aplicar la teoría de clasificación (también conocida como clasificación de patrones) en la construcción de modelos estadísticos que permitan estimar reglas de clasificación que intenten separar lo mejor posible las categorías o tipos de clases a las que una observación, que en este caso corresponden a los estudiantes, puede pertenecer. Igualmente, derivado de los modelos de clasificación construidos, es posible usar las correspondientes reglas de clasificación para predecir adecuadamente la categoría o la clase (desconocida) a la que una futura o nueva observación (i.e. nuevo estudiante) pertenecería, condicionado a su conjunto de información relevante (por ejemplo, edad, genero, estado civil, recursos económicos, tipo de colegio, orientación profesional recibida, nivel de estudios de los padres, etcétera). Ésta última tarea, conocida como predicción de clases, es de mucha importancia para las instituciones, puesto que permitiría determinar cuáles estudiantes (nuevos) tendrían alto riesgo de presentar el evento de abandono (abandonos potenciales) y, de este modo, aplicar los programas existentes o, en su defecto, diseñar políticas de intervención para tratar este conjunto de estudiantes e intervenir en su posible decisión de abandono. Además, este estudio serviría como línea de base para la evaluación de impacto de las políticas institucionales implementadas.

En este orden de ideas, el presente documento se divide en tres secciones. En

la segunda, se describe brevemente la información de la encuesta. En la tercera, se presenta el problema de la teoría de clasificación de patrones de manera intuitiva, los modelos usados y los resultados obtenidos a partir de su estimación para el caso de la Universidad de Antioquia.

2. Descripción de la encuesta

En el marco del proyecto Alfa-Guía se diseñó un formulario sobre el abandono en la educación superior, el cual fue validado en 12 países de América Latina y el Caribe, España y Portugal. Las preguntas incluidas dan cuenta de las circunstancias previas al inicio del estudiante en la IES y al abandono y los factores que lo motivan. Dichas preguntas hacen referencia a las variables asociadas, que según la revisión de la literatura son sus principales determinantes. Estos factores se clasifican en: académicos, institucionales, socio-económicos, individuales/familiares y culturales.

La encuesta se realizó telefónicamente en el 2013 por la firma española Análisis e Investigación: estudios de mercado, marketing y opinión −AEI−. El listado de estudiantes encuestados para la Universidad de Antioquia fue de corte transversal correspondiente a estudiantes matriculados en los años 2008, 2009 y 2010, en la sede central, Medellín, y en las diferentes regiones del Departamento de Antioquia en las que se tienen Sedes: Oriente (Carmen de Viboral), Occidente (Santafé de Antioquia), Magdalena Medio (Puerto Berrio), Bajo Cauca (Caucasia), Norte (Yarumal), Nordeste (Amal y Segovia), Uraba (Turbo y Apartadó), Suroeste (Andes), Sonsón y Envigado.

Además de los factores, la encuesta se estructuró en cuatro bloques de preguntas que sirven para alcanzar diferentes objetivos, como se describe a continuación.

Bloque 0 (preguntas institucionales). Recoge información características individuales (edad y sexo) y académicas básicas (nombre de la carrera en la que se matriculó, rama del conocimiento, modalidad, si continúa activo, si abandonó, calificación en la prueba de acceso a la IES y carga académica) de los estudiantes. La información fue reportada por cada IES y contiene información para 14 preguntas. Adicionalmente, incluye una clasificación inicial de los estudiantes por tipologías de abandono. Esta información sirvió para hacer control de calidad a la empresa que realizó la encuesta y para rastrear a los estudiantes que finalmente fueron encuestados.

Bloque 1 (preguntas generales). Bloque compuesto por 33 preguntas que pertenecen a cada uno de los factores que, teóricamente, determinan el abandono, las cuales corresponden a variables explicativas del fenómeno del abandono. Véase la Tabla 1.

Bloque 2 (preguntas de posicionamiento). Comprende cinco preguntas que permiten clasificar a los estudiantes en cinco posibles categorías o clases:

a) Estudiantes que continúan matriculados en la misma carrera en la que se matricularon inicialmente en la IES (Ac).
b) Estudiantes matriculados en un programa académico diferente ofrecido en la misma IES (CP).

c) Estudiantes matriculados en otra IES (CIES).
d) Estudiantes matriculados en otra institución académica de nivel no universitario (CN).
e) Estudiantes que abandonan o interrumpen de forma temporal o definitiva sus estudios (Ab).

A partir de esta clasificación, y desde el punto de vista de la IES, se pueden estimar modelos estadísticos para clasificar a los estudiantes en dos clases (modelos binomiales para activos y que abandonaron) o en cinco clases (modelos multinomiales).

Bloque 3 (preguntas por perfiles). De acuerdo a las respuestas obtenidas en el Bloque 2, se hicieron preguntas específicas para los estudiantes en cada tipo de clase, de donde se obtiene un análisis por perfil.

Tabla 1: Variables predictoras

Factores	**Variables**
Individuales/familiares:	- Estado civil - Con quién vive - Número de hermanos - Lugar que ocupa entre sus hermanos - Al menos un hermano cuenta con educación superior - Experiencias familiares (separación, muerte, entre otras) - Entorno familiar en cuanto a hábitos de estudio
Académicos	- Título más alto con el que ingresó - Tipo de institución en la que terminó el bachillerato - Interrupción de estudios en educación media - Años entre el bachillerato y el inicio en la IES - Puntuación mínima requerida en la prueba de ingreso a la IES - Motivo de la elección de la carrera - Valoración del nivel alcanzado en aspectos como: formación previa, cumplimiento de compromisos en el programa, etc. - Metodología de enseñanza - Satisfacción con diversos aspectos de la carrera
Socioeconómicos	- Nivel de estudios de los padres/tutor - Lugar de procedencia - Estado de salud - Creencias y prejuicios negativos hacia las profesiones que requieren título de educación superior - Dependencia económica - Recursos económicos - Apoyo económico (beca, crédito, subsidios, etc.)
Institucionales	- Reconocimiento como miembro de alguna minoría étnica dentro de la institución - Experiencias negativas vividas en la institución - Necesidades educativas especiales - Ambiente de convivencia en la institución - Participación en grupos - Adaptación a la vida en la IES - Orientación profesional ofrecida por la IES - Satisfacción con aspectos de gestión

En total se encuestaron 1.123 estudiantes, de los cuales el 63.58 % (714) corresponden a Medellín y el 36.42 % (409) a las regiones. De las 1.123 encuestas aplicadas sólo se obtuvieron datos confiables para 1.073, lo que representa una pérdida del 4 % de las observaciones. De esta se obtuvo información completa para 765 estudiantes, (93 % del total).

Para la sede Medellín, de los 676 estudiantes, 162 son estudiantes que continúan matriculados en la misma carrera en la que se matricularon inicialmente en la IES (AC), y los restantes 514 presentaron alguno de los siguientes casos: matriculados en un programa académico diferente ofrecido en la misma IES (CP), matriculados en otra IES (CIES), matriculados en otra institución académica de nivel no universitario (CN), o estudiantes que abandonan o interrumpen de forma temporal o definitiva sus estudios (Ab).Dentro de la categoría de abandono se presenta una mayor frecuencia de estudiantes que cambiaron de IES (42.8 %). Mientras que para las regiones, de los 397 estudiantes, 174 son estudiantes activos en el programa matriculado inicialmente en la IES, y los 223 restantes presentaron alguno de los casos enumerados anteriormente. Para las diferentes 11 sedes, la distribución de los estudiantes en las 5 diferentes posibles clases consideradas se aprecia en la Tabla 2, presentando una mayor frecuencia de estudiantes en la categoría de abandono definitivo con un porcentaje del 43.94 % .

Tabla 2: Distribución estudiantes en las 5 diferentes posibles clases por sede

Sede/Clase	**Ac**	**CP**	**CIES**	**CMN**	**Ab**	**Subtotal**[a]	**Total**
Medellín	162	87	220	26	181	514	676
Oriente	48	13	27	5	25	70	118
Occidente	14	1	5	2	9	17	31
Magdalena Medio	10	0	3	2	4	9	19
Bajo Cauca	35	6	10	5	19	40	75
Norte	8	4	9	4	6	23	31
Nordeste	9	1	2	1	6	10	19
Uraba	11	1	4	1	4	10	21
Suroeste	29	3	6	0	14	23	52
Sonsón	2	0	0	0	3	3	5
Envigado	8	3	6	1	8	18	26
Subtotal[b]	174	32	72	21	98	223	397
Total	336	119	292	47	279	737	1073

[a] Suma de las observaciones de las clases CP, CIES, CMN y AB por sede.
[b] Suma de las observaciones de las sedes regionales (i.e. sin Medellín) por clase.

3. Clasificación de patrones

Los métodos y algoritmos de la teoría de clasificación de patrones ha sido exitosamente aplicada en la clasificación de clases de tumores de cáncer, identificación de mensajes de correos *spam*, reconocimiento de objetos y rostros en imágenes, categorización de textos, entre muchas otras. El problema de clasificación se define del siguiente modo. Asúmase que se observa un conjunto de n parejas $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ independientes e idénticamente distribuidas (i.i.d.) acorde con una función de distribución de probabilidad desconocida $\mathbb{P}(\boldsymbol{x}, y)$, donde $\boldsymbol{x}_i \in \mathbb{R}^p$ es un vector de p variables predictoras (covariables), y $y \in \{1, 2\}$ una variable de respuesta binaria que indica si la i-ésima observación pertenece a una de las dos posibles categorías o clases, 1 o 2. En el caso del problema del abandono estudiantil, $y_i = 1$ categoriza a la i-ésima observación como estudiante activo y $y_i = 2$ como un estudiante que abandonó los estudios, $K = 2$. En la encuesta la variable `var2_1` categoriza estos dos posibles eventos.

Sin embargo, el hecho de que un estudiante haya abandonado el programa académico en la institución de educación superior, en este caso, la Universidad de Antioquia, no implica necesariamente que éste haya abandonado definitivamente el sistema de educación superior, lo que genera más de dos posibles categorías o clases ($K > 2$ clases), $y \in \{1, 2, \ldots, K\}$. Así, en la encuesta se indagó por las siguientes posibilidades: estar *a*) matriculado en el mismo programa académico en el que se inscribió inicialmente (`var2_1`), *b*) estar matriculado en un programa académico diferente en la misma institución en la que estaba inscrito inicialmente (`var2_2`), *c*) estar matriculado en una institución de educación superior diferente (`var2_3`), *d*) estar matriculado en una institución educativa de nivel no universitario (`var2_4`), ó *e*) no estar matriculado en ninguna institución de educación, es decir, interrumpió temporal o definitivamente los estudios (`var2_5`).

En consecuencia, dado el conjunto de entrenamiento $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, el objetivo del problema de clasificación consiste en aprender la regla de decisión óptima

$$\phi(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\arg\max}\, f_k(\boldsymbol{x})$$

que prediga con precisión las K-clases para observaciones futuras, donde las funciones $f_k \colon \mathbb{R}^p \to \mathbb{R}$ representan la fortaleza de evidencia de que una observación con vector de insumos $\boldsymbol{x}$ pertenezca a la clase k, $k = 1, \ldots, K$. Así, la función (o clasificador) ϕ asigna una observación, con vector $\boldsymbol{x}$, a la clase k con mayor función $f_k(\boldsymbol{x})$.

En la literatura existen numerosas metodologías estadísticas de clasificación de patrones para $K = 2$ ó $K > 2$ clases, por ejemplo: *a*) técnicas de análisis discriminante, *b*) modelos lineales generalizados (en el cual se encuentra el popular modelo de regresión logístico), *c*) modelos aditivos generalizados, *d*) métodos basados en árboles de decisión, *e*) redes neuronales, *f*) máquinas de soporte vectorial, etcétera. Sin embargo, la pregunta obvia que resulta ante la variedad de técnicas de modelación de clasificación de patrones es, cuál(es) de éstas emplear. La respuesta no es trivial debido a que hay que tener en cuenta varios aspectos, tales como la capacidad o potencia predictiva, la interpretabilidad de los resultados, la robustez con respecto a la presencia de datos atípicos en las variables, la dependencia sobre hiper-parámetros, la complejidad computacional, entre otros (véase Steinwart and Christmann [21] para una comparación de las propiedades de las principales técnicas de clasificación). Adicionalmente, la calidad de los datos empleados desempeña una función de mucha importancia. Por ejemplo, una calidad de datos deficiente hace que sea muy difícil justificar algunos supuestos (paramétricos) de los modelos, y una base de datos grande a menudo contiene muchos errores, diferentes tipos de datos atípicos, inconsistencias, datos perdidos (*missing data*), etcétera. Por lo tanto, el conocimiento obtenido a partir del proceso de minería de datos es indispensable.

Aunque se pudieron emplear varias técnicas de clasificación, se optó por aplicar los modelos lineales generalizados, en particular, los modelos de regresión logística de dos clases y de regresión multinomial (i.e. regresión logística con múltiples clases). La elección del modelo lineal generalizado se debió a su popularidad y a que, inicialmente, el Proyecto Alfa Guía diseñó la encuesta para aplicar un modelo de regresión logístico binario.

Sin embargo, para la correcta aplicación de la metodología de clasificación propuesta, es importante identificar las variables predictoras relevantes, esto es, cuáles de las p variables en el vector $\boldsymbol{x}$ son importantes para construir las funciones de clasificación. Este problema de selección de variables es de especial importancia, en particular cuando el vector $\boldsymbol{x}$ es de alta dimensión (i.e. el número de variables p es grande con respecto al número de observaciones n). Puesto que la inclusión de muchas variables redundantes en un modelo pueden afectar negativamente su precisión predictiva, la selección de variables es importante y necesaria para hacer clasificaciones precisas, en especial cuando el problema de clasificación es de múltiples clases ($K > 2$). Estos predictores redundantes incluyen tanto variables con errores y variables que están altamente correlacionadas con otras. En la encuesta aplicada se obtuvo información para un número considerable de variables predictoras (87 en total), la mayoría de las cuales son variables de tipo multinomial y ordinal, que a su vez requieren de la construcción de variables *dummy* con el fin de poder incluirlas en los modelos propuestos, dejando una de las categorías y/o calificación como base (se obtuvo así un total de 161 variables predictoras). Por lo tanto, es importante llevar a cabo el proceso de clasificación y selección de variables conjuntamente con el fin de tener modelos de clasificación parsimoniosos (i.e. con pocas variables relevantes) precisos y fáciles de interpretar.

En la literatura estadística existen múltiples estrategias de selección de variables en modelos de regresión y clasificación (véase por ejemplo Hastie et al. [16]). Una de las técnicas más eficientes y empleadas en la actualidad es la de regularización (o penalización), la cual consiste en imponer penalidades sobre alguna función del vector de coeficientes $\boldsymbol{\beta}$ asociado al vector de variables predictoras $\boldsymbol{x}$, con el fin de identificar cuáles coeficientes β_j asociados a las variables x_j, $j = 1, \ldots, p$ son exactamente iguales a cero, en cuyo defecto implica que la correspondiente variable es redundante para predecir la variable de respuesta y. En consecuencia, la selección de variables permite identificar aquellas variables cuyos coeficientes de regresión son diferentes de cero, i.e. $J(\boldsymbol{\beta}) = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$, basado en el supuesto de que muchos coeficientes son cero (i.e. *sparse assumption*). La estimación del vector de parámetros asociado al modelo tiene la propiedad de que éste selecciona las variables (relevantes) en $\boldsymbol{x} \in \mathbb{R}^p$ en el sentido que $\widehat{\boldsymbol{\beta}}(\lambda) = 0$ para algunos j's dependiendo de la selección del parámetro de regularización o penalización $\lambda \geq 0$, el cual determina el monto de reducción (encogimiento) del número variables, donde a mayor λ mayor restricción sobre el número de variables a incluir en el modelo. En la práctica, el parámetro λ es elegido por medio de algún método que proporcione la optimalidad de la capacidad de predicción, tal como la técnica de validación cruzada empleada en el presente informe (e.g. Hastie et al. [16]).

Después de depurar los datos por inconsistencias, la base de datos se dividió a su vez en tres bases: una para la sede Medellín, otra para las sedes regionales, y la última que agrupa la sede Medellín y las sedes de las regiones. Además, las observaciones para cada una de estas base de datos se dividieron en observaciones seleccionadas aleatoriamente (sin reemplazo) para entrenar las reglas de clasificación, correspondientes a un 70 % del total (i.e. datos de entrenamiento), y el 30 % de las observaciones restantes se utilizaron para evaluar la capacidad predictiva de los modelos estimados a través del

error de predicción y/o clasificación (i.e. datos de prueba). El número de observaciones (estudiantes) en los datos de entrenamiento y prueba para cada clase, tanto en el caso binomial (dos clases $K = 2$) como multinomial (múltiples clases, $K = 5$), están reportados en las Tablas 3 y 4.

Tabla 3: Distribución de las clases

Medellín			
Datos	**Clase 1** (Activo)	**Clase 2** (Abandono)	**Total**
Entrenamiento	114	359	473
Prueba	48	155	203
	162	514	676
Regiones			
Datos	**Clase 1** (Activo)	**Clase 2** (Abandono)	**Total**
Entrenamiento	118	159	277
Prueba	56	64	120
	174	223	397
Medellín y Regiones			
Datos	**Clase 1** (Activo)	**Clase 2** (Abandono)	**Total**
Entrenamiento	234	517	751
Prueba	102	220	322
	336	737	1073

Tabla 4: Distribución múltiples clases

Medellín						
Datos	**Clase 1** (NC)	**Clase 2** (CP)	**Clase 3** (CIES)	**Clase 4** (CMN)	**Clase 5** (Abandono)	**Total**
Entrenamiento	117	58	155	18	125	473
Prueba	45	29	65	8	56	203
	162	87	220	26	181	676
Regiones						
Datos	**Clase 1** (NC)	**Clase 2** (CP)	**Clase 3** (CIES)	**Clase 4** (CMN)	**Clase 5** (Abandono)	**Total**
Entrenamiento	119	23	53	13	69	277
Prueba	55	9	19	8	29	120
	174	32	72	21	98	397
Medellín y Regiones						
Datos	**Clase 1** (NC)	**Clase 2** (CP)	**Clase 3** (CIES)	**Clase 4** (CMN)	**Clase 5** (Abandono)	**Total**
Entrenamiento	227	85	197	35	207	751
Prueba	109	34	95	12	72	322
	336	119	292	47	279	1073

3.1. Modelos logísticos

A continuación se presentan los resultados obtenidos de la aplicación de los modelos logísticos binarios y multinomiales regularizados con el objeto de seleccionar el conjunto de variables relevantes para predecir las diferentes clases. A través de la aplicación de la técnica de validación cruzada para obtener el valor del parámetro de regularización óptimo se obtuvieron dos tipos de valores λ_{min}, correspondiente al valor de λ que correspondiente a la devianza media mínima, y λ_{1se}, correspondiente al valor más grande de λ tal que el error (devianza) esté dentro de un error estándar del valor mínimo (véase Friedman et al. [13]).

3.1.1. Modelos logísticos binarios

Como se aprecia en las Gráficas 1, 2 y 3, y en la Tabla 5, el número de variables seleccionadas (17) para los estudiantes de la sede Medellín, con base en el parámetro óptimo estimado $\lambda_{\text{1se}} = 0.032$ es menor con respecto al número seleccionado (58) con base en $\lambda_{\text{min}} = 0.014$, con un error de clasificación aproximado de 0.18 en ambos casos. Para los estudiantes de las 10 regiones se encontró que el número de variables relevantes fue 7 con un $\lambda_{\text{1se}} = 0.071$, menor a las 59 variables seleccionadas con base en $\lambda_{\text{min}} = 0.021$. Sin embargo, el error de clasificación aumentó en comparación con Medellín, esto podría estar asociado a inconsistencias y errores en los datos. Finalmente, el modelo conjunto, es decir Medellín y Regiones, arroja un total de 18 (37) variables relevantes con base en $\lambda_{\text{1se}} = 0.035$ ($\lambda_{\text{min}} = 0.022$). Par este caso el error de clasificación es de 0.30.

Tabla 5: Selección de variables y errores de predicción para los modelos logísticos

Medellín			
Parámetro de regularización λ	**log(λ)**	**No. variables seleccionadas**	**Error**
$\lambda_{\text{min}} = 0.014$	-4.271	58	0.1872
$\lambda_{\text{1se}} = 0.032$	-3.434	17	0.1823
Regiones			
Parámetro de regularización λ	**log(λ)**	**No. variables seleccionadas**	**Error**
$\lambda_{\text{min}} = 0.021$	-3.858	59	0.425
$\lambda_{\text{1se}} = 0.071$	-2.649	7	0.40
Medellín y regiones			
Parámetro de regularización λ	**log(λ)**	**No. variables seleccionadas**	**Error**
$\lambda_{\text{min}} = 0.022$	-3.812	37	0.311
$\lambda_{\text{1se}} = 0.035$	-3.346	18	0.292

Las Tablas 6-7, 8-9 y 10 listan las variables seleccionadas para los modelos logísticos binarios usando los datos de entrenamiento de las tres bases de datos (Medellín, Regiones y Medellín-Regiones) con base en los valores óptimos del parámetro de regularización λ con sus respectivos valores estimados de los parámetros de regresión.

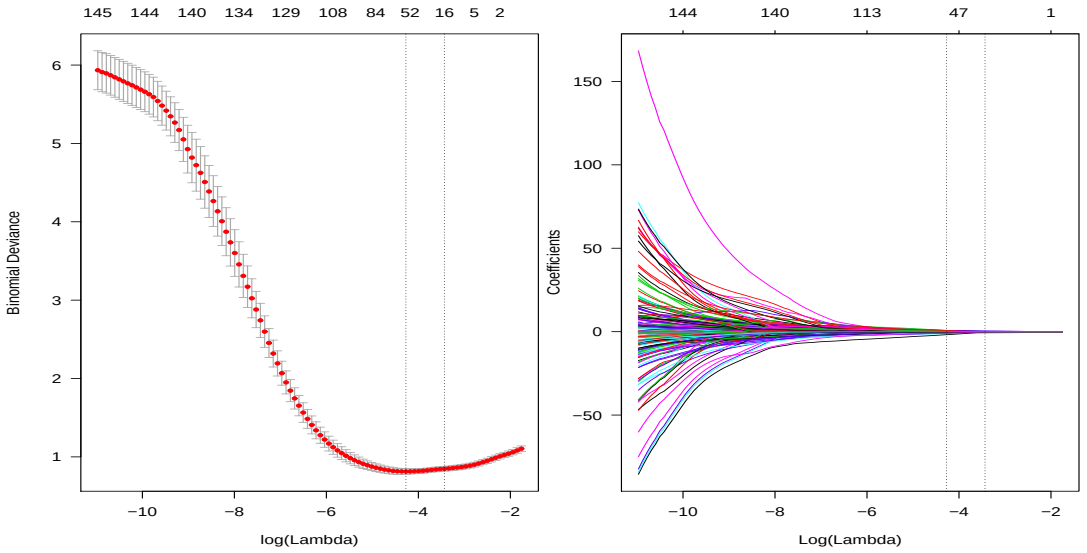

Figura 1: **Medellín**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1) estimados. La línea vertical izquierda corresponde al mínimo error, mientras que la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del valor mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

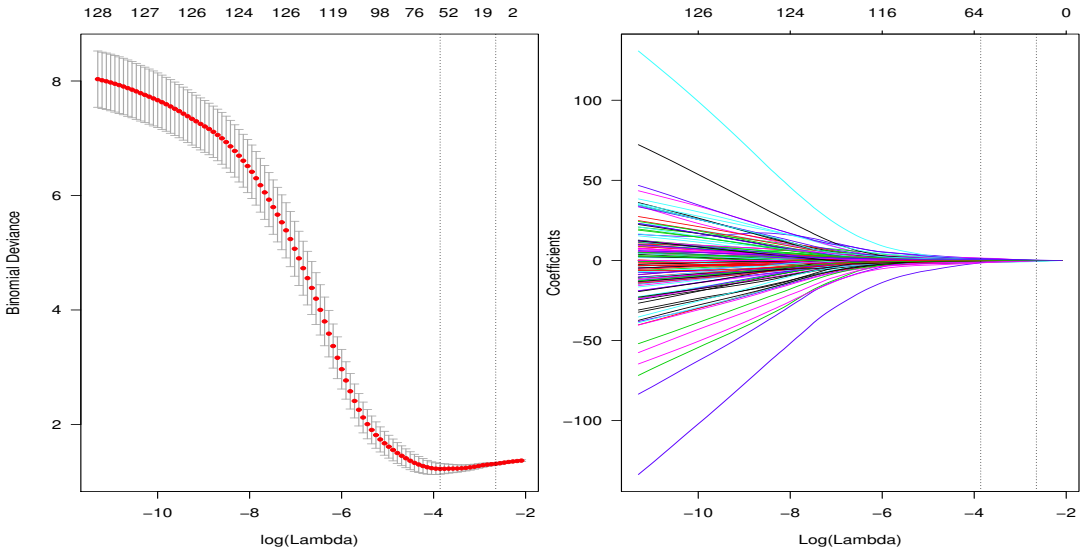

Figura 2: **Regiones**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1) estimados. La línea vertical izquierda corresponde al mínimo error, mientras que la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del valor mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

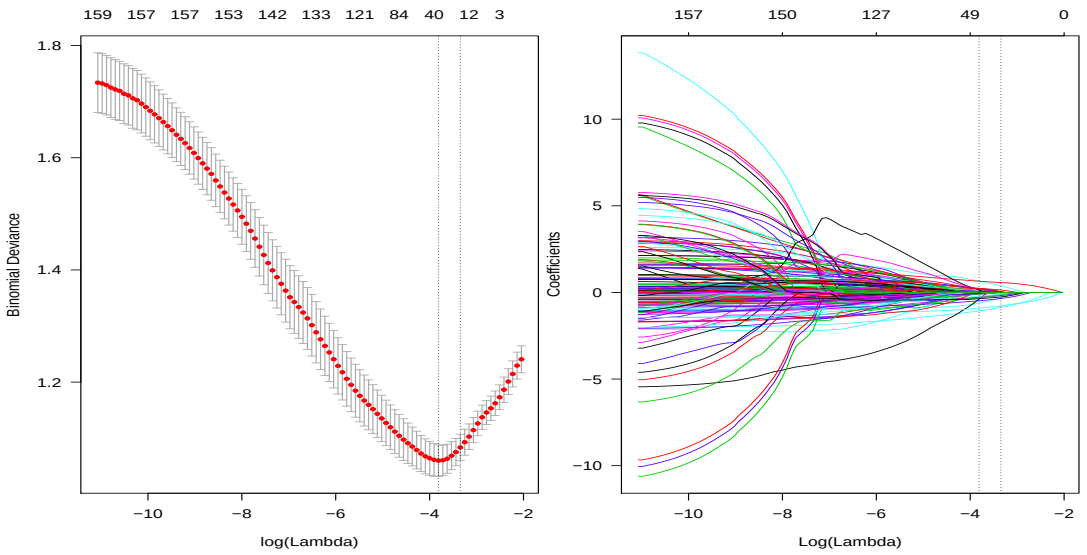

Figura 3: **Medellín y Regiones**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1) estimados. La línea vertical izquierda corresponde al mínimo error, mientras que la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del valor mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

Al comparar los resultados obtenidos por λ_{min} y λ_{1se} para la sede Medellín, se encontraron factores que podrían denominarse protectores para la permanencia, clasificados en previos a la entrada a la IES y en aquellos que se adquieren con la interacción académica e institucional. En cuanto a los previos, se puede considerar el menor riesgo de abandono en estudiantes de carreras presenciales en comparación con aquellos que cursan pregrados a distancia, recibir apoyo económico en forma de becas o subsidios o trabajo en la institución, elegir la carrera por vocación o por orientación profesional. En cuanto a la pregunta relacionada con la persona con quién vive se nota un mayor riesgo de abandono si vive con el cónyuge y menor si vive en residencia estudiantil, este último implica movilidad y vínculos regionales generados por dicha movilidad. Ahora, entre los que se obtienen después de iniciar los estudios, se destacan: tener un rendimiento académico que le permita tener mayor carga académica, estar muy satisfecho con la calidad docente y estar satisfecho con las condiciones de seguridad que ofrece la gestión institucional. De otro lado, las variables que parecen generar un mayor riesgo de abandono son: no recibir ningún tipo de ayuda económica, haber vivido alguna experiencia traumática dentro de la institución y vivir con el cónyugue.

Para el modelo regional, es importante anotar inicialmente que el error de clasificación fue de 0.40, independientemente del valor óptimo para λ. En particular, aquellos estudiantes que viven con familiares, reciben apoyo económico en forma de becas, subsidios o trabajo institucional tienen mayor riesgo de permanencia. En cuanto al incremento del riesgo de abandono se destacan la brecha entre la finalización del bachillerato y el inicio en la IES y declarar una adaptación académica regular.

Tabla 6: Coeficientes de las variables seleccionadas (Medellín)

	Variable	**Coeficiente** ($\boldsymbol{\lambda}_{\text{min}}$)	**Coeficiente** ($\boldsymbol{\lambda}_{\text{1se}}$)
	Intercepto	4.851	3.511
	20 años	-0.268	
	Género	-0.120	
	Agricultura	-0.103	
	Modalidad	-1.312	-1.049
	% de carga académica	-0.005	-0.005
Vive con	Familiares	-0.223	
	Cónyuge	0.515	0.168
	Residencia	-1.824	-0.185
Escolaridad padre	Sin escolaridad (no alfabetizado)	0.042	
Escolaridad madre	Sin escolaridad (no alfabetizado)	0.143	
	Muy bueno	0.506	0.123
	Cambio est. civil	-0.183	
	Probs. Psíquicos	0.438	
	Ingreso laboral	0.222	
	Perdida empleo	-0.741	
Dependencia eca.	Otros familiares	0.672	
	Cónyuge	0.715	
	Otra persona	-0.647	
	Suficiencia recursos ecos.	0.238	
Apoyo eco.	Becas o subsidios	-0.708	-0.503
	Trabajo en la institución	-0.232	-0.097
	No tuvo ayuda	0.730	0.496
	Interrupción temporal de estudios	-0.005	
	3 años	-0.198	
	Vocación	-0.539	-0.223
	Orientación profesional	-0.400	-0.037
	Discriminación	0.367	
	Otras experiencias	0.762	0.064
Relación con profesores	Mala	0.560	
	Regular	0.144	
	Muy buena	-0.232	
	Buena	-0.087	-0.052
	Participación política	-0.277	
	Participación académica	-0.788	-0.574
	Participación deportiva	0.007	
Adaptación social	Mala	0.122	
	Regular	0.042	
Adaptación académica	Mala	0.410	
	Buena	-0.225	
	Orientación profesional de la Inst.	-0.491	-0.137
Orientación	Insatisfecho	0.498	
	Muy satisfecho	-0.155	
Coordinación	Insatisfecho	0.087	
	Satisfecho	0.032	
	Muy satisfecho	-0.020	
	Poco satisfecho	-0.006	
	Satisfecho	0.006	
Calidad docente	Insatisfecho	0.285	
	Muy satisfecho	-0.524	-0.132
	Poco satisfecho	-0.061	
Exigencia	Poco satisfecho	-0.307	
	Muy satisfecho	0.067	

Tabla 7: Coeficientes de las variables seleccionadas (Medellín) cont.

	Variable	**Coeficiente ($\lambda_{\min}$)**	**Coeficiente (λ_{1se})**
Calidad del programa	Poco satisfecho	0.484	
	Satisfecho	0.201	
Seguridad	Satisfecho	-0.245	-0.062
Espacios físicos	Insatisfecho	-0.953	
	Poco satisfecho	-0.216	

3.1.2. Modelos multinomiales

Para el problema de clasificación con múltiples clases (i.e. con $K = 5$ clases), la Tabla 11 reporta el número de variables seleccionadas con base en los valores óptimos $\lambda_{\min}$ y λ_{1se} para las tres bases de datos consideradas. Véase también las Figuras 4, 5 y 6, donde se aprecian las gráficas de los perfiles (o trayectorias) para cada uno de los coeficientes de regresión asociados a las variables predictoras versus diferentes valores del parámetro de regularización λ, $\widehat{\boldsymbol{\beta}}(\lambda)$. Así como en los modelos logísticos de dos clases el número de variables seleccionadas con base en λ_{1se} es menor con respecto a las variables relevantes con base en $\lambda_{\min}$. Adicionalmente, como es natural, debido a que se tienen más clases, el error de clasificación (0.47) con base en los datos de prueba es mayor que con respecto al caso de sólo dos clases.

En las Tablas 12-13, 14-15-16 y 17-18-19 se listan las variables seleccionadas y sus correspondientes coeficientes de regresión estimados con base en las tres bases de datos empleadas (Medellín, Regiones y Medellín-Regiones) con los parámetros de regularización $\lambda_{\min}$ y λ_{1se} para cada una de las cinco clases. Pare el modelo de Medellín se obtuvieron resultados de variables explicativas relevantes, diferentes a las encontradas en los modelos binomiales tales como el género, la educación de los padres y la convivencia. Particularmente se destacan factores protectores para la permanecer activo como: la modalidad, el porcentaje de carga académica, las becas y subsidios y la participación académica, y como factor de riesgo el no recibir ningún tipo de apoyo económico. El cambio de programa en la misma institución estuvo motivado solamente por pertenecer al área de ciencias. El cambio de IES está influenciado por los niveles altos de la educación del padre, vivir experiencias traumáticas dentro de la institución y calificar como regulares los niveles de convivencia.

Para el modelo multinomial con base los datos de las regiones, se nota un mayor porcentaje de estudiantes clasificados en la categoría de abandono definitivo esto podría estar explicado por las pocas posibilidades que tienen algunos municipios del Departamento de educación superior, también se obtuvo en términos porcentuales una mayor frecuencia de estudiantes que se cambian a niveles de educación inferiores en comparación con Medellín.

Tabla 8: Coeficientes de las variables seleccionadas (Regiones)

	Variable	**Coeficiente** ($\boldsymbol{\lambda}_{\text{min}}$)	**Coeficiente** ($\boldsymbol{\lambda}_{\text{1se}}$)
	Intercepto	2.408	0.333
	22 años	-0.448	
	Género	-0.356	
	Ing., Indus. y Const.	-0.589	
Estado civil	Casado	0.404	
	Unión libre	0.849	
Vive con	Familiares	-1.213	-0.473
	Amigos	0.700	
	Residencia	1.343	
	Solo	0.121	
Escolaridad padre	Sin escolaridad (no alfabetizado)	0.247	
	Sin escolaridad (alfabetizado)	0.071	
	Secundaria	-0.018	
	Formación prof. superior	0.122	
Escolaridad madre	Sin escolaridad (no alfabetizado)	1.468	
	Formación prof. superior	0.037	
	No sabe	-1.025	
	Minoría étnica	-1.137	
	Muy bueno	0.312	
	Probs. Psíquicos	-0.114	
	Ingreso laboral	0.252	
	Crencias negativas a ES	0.313	
Apoyo eco.	Becas o subsidios	-1.185	-0.693
	Trabajo en la institución	-1.214	-0.275
	No tuvo ayuda	0.222	
Tiempo de inicio en la IES	1 año	0.061	
	2 años	0.013	
	3 años	0.675	0.062
	$>$ 5 años	-0.608	
	Vocación	-0.061	
	Mercado laboral	0.113	
	Acoso	-0.094	
	Discriminación	0.388	
Relación con profesores	Mala	-1.463	
	Buena	-0.014	
	Muy buena	0.062	
	Buena	-0.344	
	Participación política	0.044	
	Participación religiosa	-1.560	
Adaptación académica	Mala	0.242	
	Regular	0.838	0.033
	Muy satisfecho	-0.243	
	Poco satisfecho	-0.158	
	Satisfecho	0.045	
	Muy satisfecho	-0.059	
Calidad docente	Insatisfecho	-0.749	
	Satisfecho	0.321	
	Satisfecho	-0.192	
	Muy satisfecho	0.088	
Evaluación	Insatisfecho	0.670	
	Satisfecho	0.025	
	Muy satisfecho	-0.105	
Exigencia	Insatisfecho	-0.251	
	Poco satisfecho	-0.020	

Tabla 9: Coeficientes de las variables seleccionadas (Regiones) cont.

	Variable	**Coeficiente** ($\boldsymbol{\lambda}_{\text{min}}$)	**Coeficiente** ($\boldsymbol{\lambda}_{\text{1se}}$)
Seguridad	Insatisfecho	1.068	
	Poco satisfecho	0.324	
	Muy satisfecho	-0.245	
	Poco satisfecho	-0.263	
	Satisfecho	0.557	0.220

Tabla 10: Coeficientes de las variables seleccionadas (Medellín-Regiones)

	Variable	**Coeficiente** ($\boldsymbol{\lambda}_{\text{min}}$)	**Coeficiente** ($\boldsymbol{\lambda}_{\text{1se}}$)
	Intercepto	2.149	1.957
Edad	18 años	-0.023	
	22 años	-0.612	-0.312
	Género	-0.280	-0.182
	Modalidad	-1.065	-0.862
	Cónyuge	0.657	0.387
	Residencia	-0.385	
	Secundaria	-0.042	
	Formación prof. superior	0.031	
	Secundaria	-0.052	
	Formación prof. superior	0.022	
	Ingreso laboral	0.044	
	De sí mismo	0.128	0.002
Apoyo eco.	Becas o subsidios	-0.908	-0.780
	Trabajo en la institución	-0.263	-0.111
	No tuvo ayuda	0.647	0.568
	Tipo de colegio	0.238	0.035
	2 años	-0.005	
	Vocación	-0.222	-0.100
	Otras experiencias	0.203	
	Muy buena	-0.055	
	Regular	0.014	
	Buena	-0.038	
	Participación académica	-0.480	-0.348
	Regular	0.201	0.038
	Muy buena	-0.105	-0.024
Adaptación académica	Mala	0.144	
	Regular	0.237	0.140
Orientación	Insatisfecho	0.194	
	Muy satisfecho	-0.304	-0.155
	Satisfecho	0.047	
	Poco satisfecho	0.034	
	Poco satisfecho	-0.099	
	Satisfecho	0.065	
Seguridad	Insatisfecho	0.211	0.046
	Satisfecho	-0.048	
	Poco satisfecho	-0.362	-0.262

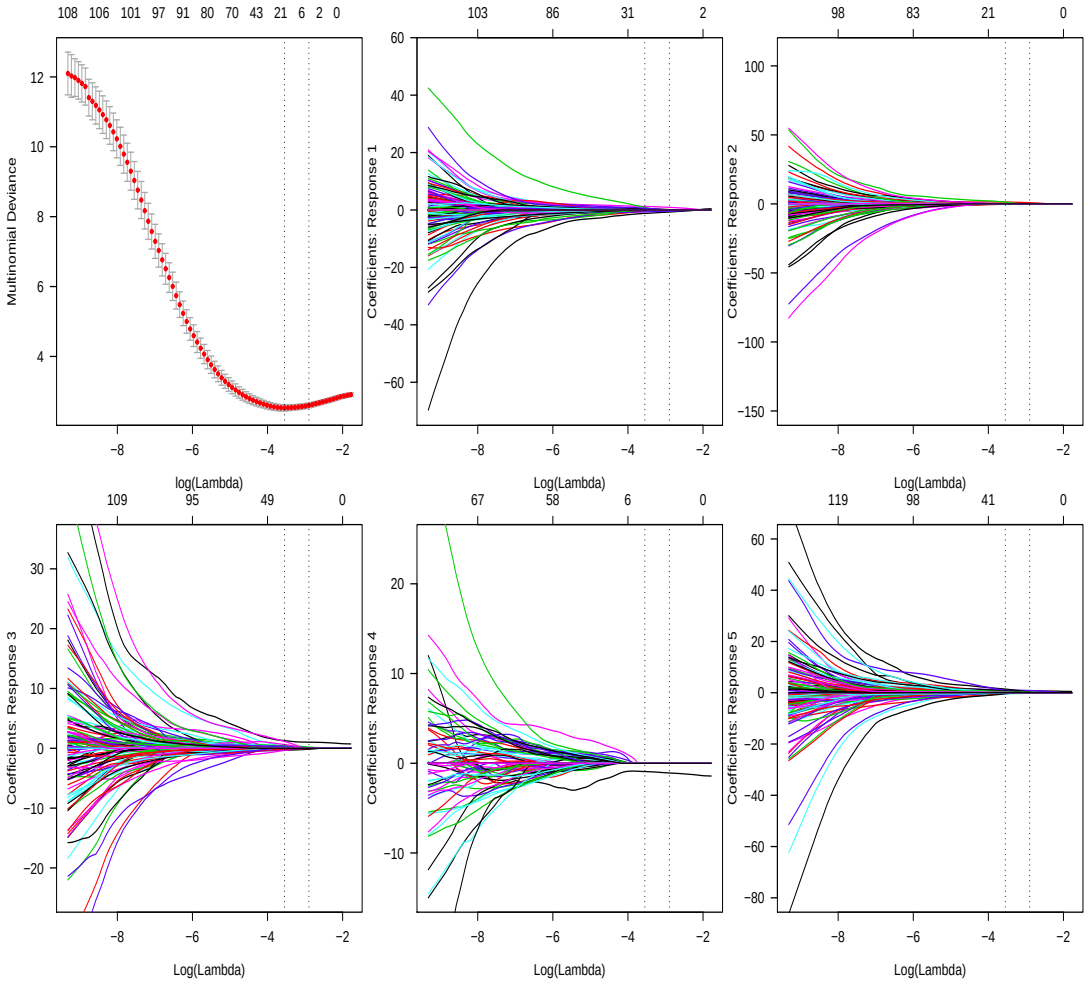

Figura 4: **Medellín**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1). La línea vertical izquierda corresponde al mínimo error, y la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

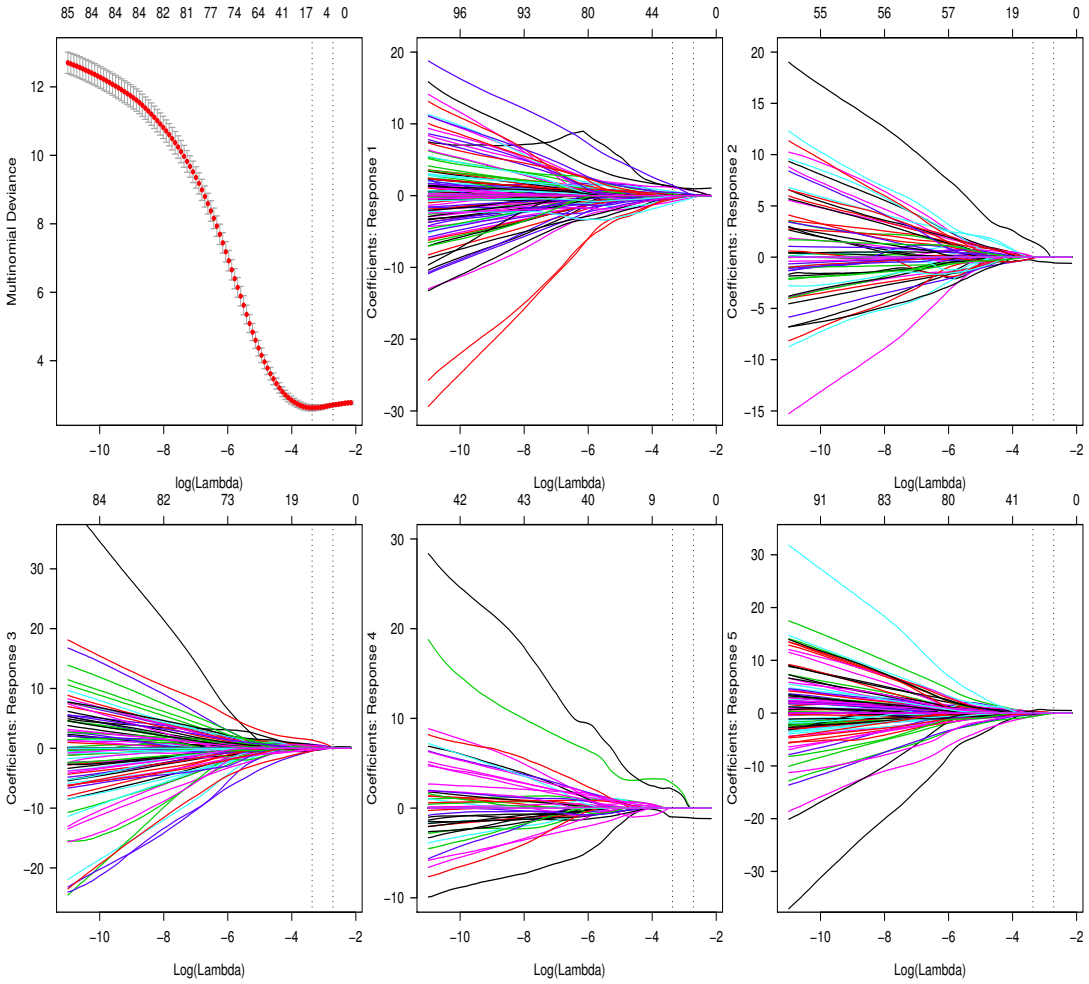

Figura 5: **Regiones**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1). La línea vertical izquierda corresponde al mínimo error, y la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

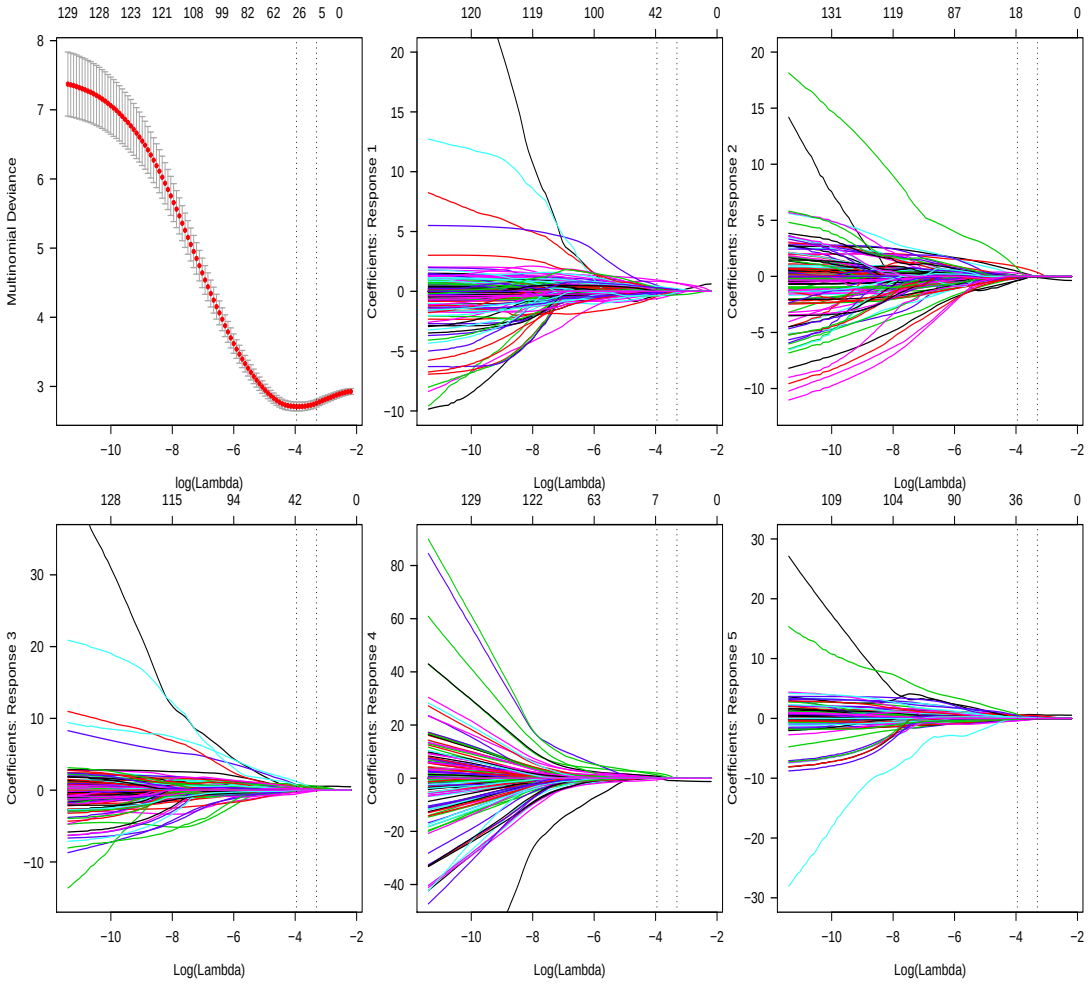

Figura 6: **Medellín-Regiones**. Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización l_1). La línea vertical izquierda corresponde al mínimo error, y la linea vertical derecha corresponde al mayor valor de λ tal que el error esté dentro de un error estándar del mínimo. En la parte superior de las gráficas se especifica el tamaño del modelo

Tabla 11: Selección de variables y errores de predicción para los modelos multinomiales

Medellín							
$\boldsymbol{\lambda}$	$\log(\boldsymbol{\lambda})$	**No. variables seleccionadas**					**Error**
		Cl. 1	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**	
$\lambda_{\min} = 0.0287$	-3.549	19	7	24	1	22	0.507
$\lambda_{1se} = 0.0551$	-2.898	6	2	7	1	7	0.498
Regiones							
$\boldsymbol{\lambda}$	$\log(\boldsymbol{\lambda})$	**No. variables seleccionadas**					**Error**
		Cl. 1	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**	
$\lambda_{\min} = 0.0345$	-3.365	28	6	17	4	23	0.5
$\lambda_{1se} = 0.0662$	-2.714	7	1	4	1	4	0.542
Medellín y regiones							
$\boldsymbol{\lambda}$	$\log(\boldsymbol{\lambda})$	**No. variables seleccionadas**					**Error**
		Cl. 1	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**	
$\lambda_{\min} = 0.019$	-3.957	41	14	37	7	29	0.531
$\lambda_{1se} = 0.037$	-3.305	15	2	12	1	9	0.525

4. Conclusiones

Después de realizar el análisis de las variables relevantes obtenidas de los modelos binomiales y multinomiales, ya sea para dos o cinco clases, se podría concluir que existen unas variables que pudieran denominarse transversales, las cuales son comunes a las Regiones y a Medellín, y explican la clasificación de los estudiantes en cada clase, independientemente de la sede. Estas variables son: edad, apoyo económico y adaptación académica, encontrando un menor riesgo de abandono en estudiantes de 22 años, que reciben becas o subsidios o tienen algún trabajo dentro de la institución y manifiestan tener una buena orientación académica.

Tabla 12: Coeficientes de las variables seleccionadas con base en $\lambda_{\min}$ (Medellín)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Intercepto	-1.182	0.038	1.277	-0.926	0.793
Edad	21 años					0.121
	22 años					0.253
	25 años			0.168		
	$>$ 25 años					1.093
Área conocimiento	Ciencias		1.335			
	Modalidad	1.238				
	% de carga académica	0.005				
Estado civil	Viudo			0.130		
Vive con	Cónyuge	-0.158				0.373
	Residencia	0.887				
Escolaridad padre	Secundaria			-0.325		
	Posgrado			0.295		
	No sabe			0.293		
Escolaridad madre	Secundaria			-0.157		
	Universitaria			0.182		
	No sabe			0.009		
Procedencia	Otro país con residencia por estudios		1.333			
Estado salud	Malo					1.436
	Muy bueno			0.427		
Experiencia	Cambio est. civil			-0.354		
	Probs. Psíquicos	-0.103				
	Ingreso laboral					0.471
	Perdida empleo	0.395				
Dependencia eca.	Otros familiares	-0.308				-0.256
	Cónyuge					0.080
	Suficiencia recursos ecos.					0.184
Apoyo eco.	Becas o subsidios	0.049		-0.169		
	No tuvo ayuda	-0.654				
	Tipo de colegio					-0.442
Tiempo de inicio en la IES	2 años		-0.226			
	3 años	0.035				-0.113
	5 años					-0.054
	$>$ 5 años			-0.250		
	Vocación	0.130		-0.144		
	Discriminación		0.130			-0.225
	Otras experiencias			0.337		
Relación con profesores	Mala			1.138		
	Regular	-0.044				
Relación con compañeros	Mala					0.880
	Buena			-0.098		
	Muy buena	0.021				
Convivencia	Regular			0.238		
	Muy buena		0.110			
	Participación académica	0.468				
	Participación cultural	0.053		-0.065		
Adaptación social	Mala			0.635		
	Regular					0.040
	Orientación profesional de la Inst.			-0.109		
Orientación	Poco satisfecho			0.199		
	Muy satisfecho	0.071				
Coordinación	Satisfecho					0.006
Atención del profesor	Insatisfecho					-0.405
	Poco satisfecho			0.074		
	Muy satisfecho	0.010				
Exigencia	Insatisfecho					0.680
	Muy satisfecho			0.180		
Calidad del programa	Insatisfecho					0.060
	Poco satisfecho		0.019			
Seguridad	Muy satisfecho					0.438
Espacios físicos	Insatisfecho	0.604				
	Satisfecho					0.180

Tabla 13: Coeficientes de las variables seleccionadas con base en λ_{1se} (Medellín)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Intercepto	-0.712	-0.027	1.020	-1.087	0.805
Edad	$>$ 25 años					0.882
Área conocimiento	Ciencias		0.868			
	Modalidad	0.921				
	% de carga académica	0.004				
Vive con	Cónyuge					0.180
Escolaridad padre	Secundaria			-0.158		
	Posgrado			0.007		
Escolaridad madre	Secundaria			-0.031		
Estado salud	Malo					0.153
	Muy bueno			0.214		
	Ingreso laboral					0.197
	Suficiencia recursos ecos.					0.016
Apoyo eco.	Becas o subsidios	0.011				
	No tuvo ayuda	-0.419				
	Tipo de colegio					-0.145
	Otras experiencias			0.049		
Convivencia	Regular			0.014		
	Participación académica	0.139				

Tabla 14: Coeficientes de las variables seleccionadas con base en λ_{min} (Regiones)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Intercepto	1.172	-0.303	-0.396	-0.981	0.508
Edad	18 años			0.276		
	19 años			0.193		-0.396
	22 años	0.185		-0.036		
Área conocimiento	Hds. y artes					-0.547
	C.S, Ccio. y Dcho.					0.004
	Ciencias		0.034			
	Ing., Indus. y Const.	0.038				
Estado civil	Divorciado				2.731	
Vive con	Cónyuge	-0.805				
	Amigos		1.499			
	Residencia			1.364		
	Solo					0.099
Escolaridad padre	Sin escolaridad (no alfabetizado)					0.063
	Secundaria					-0.036
	Formación prof. superior					0.124
	Universitaria	-0.131		0.057		
	Posgrado					-0.662
Escolaridad madre	Sin escolaridad (alfabetizado)					0.158
Procedencia	Otro país con residencia por estudios					0.175
Experiencia	Divorcio			-0.405		
	Ingreso laboral					0.335
	Otros eventos	0.027				-0.236
	Crencias negativas a ES	-0.086				
Dependencia eca.	Cónyuge			-0.461		
	De sí mismo		-0.078			
	Suficiencia recursos ecos.					0.164
Apoyo eco.	Becas o subsidios	1.124				
	Trabajo en la institución	1.281				
	Otro					0.251
	Tipo de colegio			0.657		

Tabla 15: Coeficientes de las variables seleccionadas con base en $\lambda_{\min}$ (Regiones) cont.

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
Tiempo de inicio en la IES	3 años			0.448		
	$>$ 5 años	0.589				
	Vocación			-0.295		
	Mercado laboral					0.049
	Otro motivo					-0.017
	Discriminación	-0.208				
Relación con profesores	Regular			0.187		
	Buena	0.0005				
Convivencia	Mala				2.102	
	Muy buena					0.201
	Participación política	-0.537				
	Participación académica		0.057			-0.066
	Participación religiosa	1.442				
Adaptación social	Mala				1.94E-14	
	Buena					0.184
Adaptación académica	Regular	-0.379				
	Orientación profesional de la Inst.					-0.168
Orientación	Insatisfecho	-0.470				
	Muy satisfecho	0.101				
Coordinación	Poco satisfecho	0.218				
	Muy satisfecho			0.019		
Contenido programa	Satisfecho			-0.001		
	Muy satisfecho			0.592		
Atención del profesor	Insatisfecho	-0.300		0.008		
	Satisfecho	0.154				
Calidad de materiales	Insatisfecho					-0.103
Evaluación	Insatisfecho	-0.048				
	Poco satisfecho	0.213				
Exigencia	Poco satisfecho	-0.047				
	Satisfecho		-0.114			
Ambiente social	Poco satisfecho					-0.038
	Muy satisfecho	-0.145				
Seguridad	Insatisfecho	-1.103				
	Poco satisfecho	-0.209		0.503		
Espacios físicos	Poco satisfecho	0.458				
	Satisfecho	-0.037				

Tabla 16: Coeficientes de las variables seleccionadas con base en $\lambda_{1\text{se}}$ (Regiones)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Intercepto	0.963	-0.553	0.184	-1.1	0.530
Edad	19 años					-0.030
Vive con	Cónyuge	-0.245				
	Residencia			0.043		
Experiencia	Ingreso laboral					0.048
Apoyo eco.	Becas o subsidios	0.670				
	Trabajo en la institución	0.697				
	Otro					0.018
	Tipo de colegio			0.085		
Tiempo de inicio en la IES	$>$ 5 años	0.120				
Contenido programa	Muy satisfecho			0.039		
Seguridad	Insatisfecho	-0.245				
Espacios físicos	Poco satisfecho	0.170				

Tabla 17: Coeficientes variables seleccionadas con base en $\lambda_{\min}$ (Medellín-Regiones)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Intercepto	-0.342	-0.369	0.119	0.096	0.497
Edad	19 años			0.450		
	20 años	0.245				
	22 años	0.189		-0.523		
	23 años					0.200
	24 años					0.200
	$>$ 25 años		-0.207			0.430
	Género	0.246				
Área conocimiento	Hds. y artes	-0.010				
	C.S, Ccio. y Dcho.		0.454			0.090
	Ciencias		0.860			
	Salud y Ser. Sociales			-0.051		
	Modalidad	0.943				
Estado civil	Casado	-0.009	-0.382	0.133		
	Unión libre					0.016
	Divorciado				1.904	
	Viudo			0.739		
Vive con	Cónyuge	-0.606			0.176	
	Residencia			0.723		
Escolaridad padre	Secundaria			-0.199		
	Formación prof. superior	-0.145				0.024
	Universitaria			0.047		
	Posgrado			0.337		-0.068
	No sabe	0.087				
Escolaridad madre	Sin escolaridad (no alfabetizado)	-0.105				
	Formación prof. superior	-0.026				
	Universitaria			0.783		
	No sabe			0.114		
	Minoría étnica				-0.594	
Procedencia	Desplazado de otro país					0.208
Estado salud	Malo				0.159	
	Muy bueno			0.380		
Experiencia	Divorcio			-0.023		
	Ser madre o padre	-0.178				0.143
	Ingreso laboral					0.605
	Perdida empleo	0.153				
	Crencias negativas a ES	-0.125				
	Favorecimiento habitos est.			-0.045		
Dependencia eca.	Otros familiares					-0.048
	Cónyuge			-0.196		
	De sí mismo		-0.146			0.120
	Suficiencia recursos ecos.					0.195
Apoyo eco.	Becas o subsidios	0.848				
	Trabajo en la institución	0.306	0.060			-0.128
	Otro		-0.156			0.029
	No tuvo ayuda	-0.371				
	Tipo de colegio	-0.204	0.058			-0.371
	Interrupción temporal de estudios			0.079		
Tiempo de inicio en la IES	1 año			0.092		
	4 años	0.048				
	5 años			0.060		
	$>$ 5 años	0.181		-0.099		
	Vocación	0.279		-0.109		
	Tradición familiar					0.325
	Otro motivo					-0.165
	Discriminación				0.734	
	Otras experiencias			0.014		
Relación con profesores	Mala			1.205		
	Regular			0.189		
	Buena		0.082			
	Muy buena	0.033				

Tabla 18: Variables seleccionadas con base en $\lambda_{\min}$ (Medellín-Regiones) cont.

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
Relación con compañeros	Mala					0.661
	Buena		0.041			
	Muy buena			0.093		
Convivencia	Regular		-0.004	0.567		
	Buena	0.038				
	Participación política	-0.029				
	Participación académica	0.062	0.288	-0.327		-0.402
	Participación cultural	0.130		-0.047		-0.008
	Participación religiosa	0.494				
Adaptación social	Mala			0.676		
	Regular			0.033		
	Muy buena	0.228				-0.070
Adaptación académica	Mala		0.726		1.271	
	Regular	-0.034				
Orientación	Insatisfecho	-0.427				
	Poco satisfecho			0.533		
	Muy satisfecho	0.421				
Coordinación	Satisfecho					0.045
Calidad docente	Satisfecho	-0.118				
	Muy satisfecho	0.009				
Atención del profesor	Insatisfecho			0.315		
Calidad de materiales	Poco satisfecho	0.217				
	Muy satisfecho					0.097
	Muy satisfecho			0.237		
Calidad del programa	Insatisfecho					0.443
	Muy satisfecho	-0.063				
Seguridad	Insatisfecho	-0.378				0.155
	Poco satisfecho			0.251		
	Satisfecho	0.004		-0.001		
	Muy satisfecho	0.235				
Espacios físicos	Insatisfecho	0.279				-0.155
	Poco satisfecho	0.292				
	Satisfecho			-0.002		0.180
	Muy satisfecho			0.060		

Referencias

[1] J. P. Bean. Student attrition, intensions and confidence. *Research in Higher Education*, 17:291–320, 1980.

[2] A. Booth and S. Satchell. The hazards of doing a phd: An analysis of completion and withdrawal rates of british phd students in the 1980s. *Journal of the Royal Statistical Society*, A158:297–318, 1995.

[3] A. Cabrera, A. Nora, and M. Castañeda. Collage persistence: Structural equations modeling tests of an integrated models student retention. *The Journal of Human Resources*, 64:123–139, 1993.

[4] S. Cameron and J. Heckman. Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of american males. *The Journal of Political Economy*, 106:262–333, 1998.

[5] S. Cameron and C. Taber. Estimation of education borrowing constraint using returns schooling. Technical Report W7761, NBER Working Paper, 2001.

[6] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez. Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de Economía*, 60:

Tabla 19: Coeficientes variables seleccionadas con base en λ_{1se} (Medellín-Regiones)

	Variable	**Cl. 1**	**Cl. 2**	**Cl. 3**	**Cl. 4**	**Cl. 5**
	Variable	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5
	Intercepto	-0.111	-0.089	0.535	-0.940	0.604
Edad	19 años			0.166		
	22 años	0.015		-0.070		
	$>$ 25 años					0.290
	Género	0.090				
Área conocimiento	Ciencias		0.356			
	Modalidad	0.682				
Vive con	Cónyuge	-0.214				
Escolaridad padre	Secundaria			-0.081		
	Posgrado			0.122		
Escolaridad madre	Universitaria			0.536		
Estado salud	Muy bueno			0.211		
Experiencia	Ser madre o padre					0.056
	Ingreso laboral					0.446
Dependencia eca.	De sí mismo					0.099
	Suficiencia recursos ecos.					0.049
Apoyo eco.	Becas o subsidios	0.694				
	Trabajo en la institución	0.115				
	No tuvo ayuda	-0.293				
	Tipo de colegio					-0.100
	Vocación	0.153				
Relación con profesores	Regular			0.064		
Convivencia	Regular			0.318		
	Participación académica	0.126		-0.005		-0.051
Adaptación social	Muy buena	0.098				
Orientación	Poco satisfecho			0.347		
	Muy satisfecho	0.210				
Seguridad	Insatisfecho	-0.108				
	Poco satisfecho			0.073		
	Muy satisfecho	0.075				
Espacios físicos	Poco satisfecho	0.124				
	Satisfecho					0.092

39–66, 2004.

[7] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez. Análisis de los factores asociados a la deserción y graduación estudiantil universitaria. *Lecturas de Economía*, 65:9–36, 2006.

[8] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez. Análisis de los factores asociados a la deserción en la educación superior: un estudio de caso. *Revista de Educación*, 345:255–280, 2008. Madrid.

[9] E. Castaño, S. Gallón, K. Gómez, J. Vásquez, C. Guzmán, D. Durán, and J. Franco. *Deserción estudiantil en la educación superior colombiana: Metodología de seguimiento, diagnóstico y elementos para su prevención*. Imprenta Nacional de Colombia, 2010. Viceministerio de Educación Superior, Ministerio de Educación Nacional, Bogotá.

[10] C. Cornwell. The enrollment effects of merit-based financial aid: Evidence from Georgia's hope scholarship. Technical report, University of Georgia, Department of Economics, 2002.

[11] S. DesJardins, D. Ahlburg, and B. McCall. Simulating the longitudinal effects of ghanges in financial aid on student departure from college. *Journal of Human Resources*, 37:653–679, 2001.

[12] S. DesJardins, D. Ahlburg, and B. McCall. A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education*, 73:555–581, 2002.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01.

[14] P. Giovagnoli. Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración. Technical Report 37, Universidad Nacional de la Plata, 2002.

[15] L. Häkkinen and R. Uusitalo. The effect of a student aid reform on graduation: A duration analysis. Technical Report Working Paper Series, 8, Department of Economics, Uppsala University, 2003.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 5th edition, 2009.

[17] E. Pascarella and P. Terenzini. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, 51: 60–75, 1980.

[18] A. Porto and A. Di Gresia. Rendimiento de estudiantes universitarios y sus determinantes. Technical report, Asociación Argentina de Economía Política, 2001.

[19] S. Ruthaychonnee. Why are there dropouts among university students? experiences in a Thai university. *International Journal of Educational Development*, 32: 283–289, 2012.

[20] W. Spady. Dropout from higher education: An interdisciplinary review and synthesis. *Interchange*, 1:64–85, 1970.

[21] I. Steinwart and A. . Christmann. *Support Vector Machines*. Springer, New York,

2008.

[22] V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Education Research*, 45:89–125, 1975.

[23] J. B. Willett and J. D. Singer. From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, 61:407–450, 1991.

5. Apéndice

Tabla 20: Estadística descriptiva

Variables	**Categorías**	**Ac** $n = 188$	**CP** $n = 110$	**CIES** $n = 269$	**CN** $n = 33$	**Ab** $n = 223$	**Total** 823	$p-$**valor**
Sexo	Masculino	93	65	167	23	138	486	0.034
	Femenino	95	45	102	10	85	337	
Edad	< 18	76	54	107	9	55	301	< 0.001
	18	30	21	36	3	26	116	
	19	16	7	31	2	18	74	
	20	20	15	24	3	14	76	
	21	9	3	17	2	17	48	
	22	9	2	3	3	12	29	
	23	6	1	11	2	7	27	
	24	1	3	7	0	11	22	
	25	6	0	10	0	6	22	
	> 25	15	4	23	9	57	108	
Estado civil	Soltero	175	103	236	23	172	709	< 0.001
	Casado	6	1	19	7	25	58	
	Unión Libre	5	5	13	2	24	49	
	Divorciado	2	1	0	1	1	5	
	Viudo	0	0	1	0	0	1	
	No contesta	0	0	0	0	1	1	
Modalidad	Presencial	112	98	245	28	199	682	< 0.001
	Semipresencial	76	12	24	5	24	141	
Rama conoc.	Educación	28	7	51	8	35	129	< 0.001
	Hds. y Artes	16	8	19	2	24	69	
	C.S, Ccio. y Dcho.	38	35	54	4	53	184	
	Ciencias	16	25	15	3	13	72	
	Ing., Indus. y Const.	46	23	93	9	66	237	
	Agricultura	16	3	8	4	5	36	
	Salud y Ser. Sociales	28	9	29	3	27	96	
Con quién vive	Padres	133	85	182	18	131	549	< 0.001
	Familiares	25	12	32	3	21	93	
	Cónyuge	9	5	30	9	46	99	
	Amigos	6	1	6	0	3	16	
	Residencia	3	0	3	0	0	6	
	Solo	12	7	16	2	21	58	
	No contesta	0	0	0	1	1	2	
Hermanos con ES	Si	86	52	154	18	121	431	0.013
	No	84	43	74	9	80	290	
	No Aplica	18	15	41	6	20	100	
	No contesta	0	0	0	0	2	2	
Educ. madre	No sabe	0	1	2	1	0	4	< 0.001
	Sin escolaridad	0	0	0	0	1	1	
	Sin escolaridad	1	0	3	1	4	9	
	Primaria	52	31	52	7	73	215	
	Secundaria	86	58	110	18	101	373	
	Profesional	23	8	49	1	20	101	
	Universitaria	21	11	39	4	18	93	
	Posgrado	4	1	14	1	5	25	
	No contesta	1	0	0	0	1	2	
Minoría étnica	Si	8	6	8	5	10	37	0.031
	No	180	103	261	28	213	785	
	No contesta	0	1	0	0	0	1	
Estado salud	Muy malo	1	1	2	0	1	5	0.013
	Malo	0	0	2	2	4	8	
	Regular	16	7	13	2	17	55	
	Bueno	108	68	128	14	101	419	
	Muy Bueno	63	34	124	15	100	336	
Dependencia eco.	Padres	135	83	178	17	112	525	< 0.001
	Familiares	2	5	15	1	6	29	
	Cónyuge	2	2	2	1	8	15	
	Otra Persona	5	1	2	0	2	10	
	De sí mismo	43	19	72	14	95	243	
	No contesta	1	0	0	0	0	1	

Tabla 21: Estadística descriptiva (continuación)

Variables	**Categorías**	**Ac** $n = 188$	**CP** $n = 110$	**CIES** $n = 269$	**CN** $n = 33$	**Ab** $n = 223$	**Total** 823	$\boldsymbol{p-}$**valor**
Suficientes recursos?	Si	141	82	206	26	138	593	0.018
	No	47	28	62	7	85	229	
	No contesta	0	0	1	0	0	1	
Otros recursos	Créditos	23	18	25	3	13	82	< 0.001
	Becas/Sub	51	22	18	0	18	109	
	Trabajo IES	21	9	8	1	5	44	
	Otro	36	11	42	4	33	126	
	Ninguno	57	50	176	25	154	462	
Título de ingreso	Bachillerato	145	90	186	15	139	575	< 0.001
	Técnico/T/FP	38	19	71	14	59	201	
	Universitario	3	1	10	4	23	41	
	Otro	1	0	1	0	1	3	
	No Contesta	1	0	1	0	1	3	
Tiempo inicio en la IES	<1	84	59	134	10	87	374	0.034
	Un año	26	20	45	4	37	132	
	Dos años	28	12	37	5	35	117	
	Tres años	20	10	21	3	12	66	
	Cuatro años	8	2	7	2	9	28	
	Cinco años	6	1	11	3	10	31	
	> 5	16	6	14	6	33	75	
Vocación	Si	136	65	136	15	134	486	< 0.001
Discriminación	Si	1	4	5	3	2	15	0.006
Otras ET	Si	6	5	25	2	12	50	0.082
Ambiente IE	Muy malo	3	0	0	0	1	4	0.001
	Malo	0	0	3	1	2	6	
	Regular	14	10	57	5	23	109	
	Bueno	123	59	139	18	129	468	
	Muy Bueno	48	41	70	9	68	236	