



## **APLICACIÓN DE LA TEORÍA DE CLASIFICACIÓN AL PROBLEMA DEL ABANDONO ESTUDIANTIL: UN ESTUDIO DE CASO**

Línea Temática 1. Factores asociados al abandono. Tipos y perfiles de abandono:  
Tipo de comunicación: derivada de investigación):

GALLÓN GÓMEZ, Santiago  
VASQUEZ VELÁSQUEZ, Johanna  
Universidad de Antioquia - Colombia  
e-mail: jovasve@gmail.com

### **Resumen.**

Las causas y consecuencias socio-económicas del abandono estudiantil universitario han sido ampliamente estudiadas en la literatura académica nacional e internacional y, aunque muchas aproximaciones conceptuales han surgido para explicar por qué un estudiante abandona sus estudios a lo largo de su ciclo académico Spady (1970), Tinto (1975), Bean (1980), Cabrera et al. (1993), Pascarella (1980) y Cornwell (2002), no se dispone de aproximaciones empíricas que permitan su correcta clasificación y predicción, lo que dificulta, entre otras, la utilización eficiente de los recursos destinados para la permanencia de los mismos. Este planteamiento permite reconocer que la medición y predicción del riesgo de abandono es fundamental para la toma de decisiones y un insumo necesario en la evaluación de impacto de las políticas institucionales implementadas. Así, y con el objetivo de desarrollar una metodología que cumpla con tal propósito, en este trabajo se estudian los determinantes del abandono estudiantil universitario aplicando la teoría de clasificación de patrones en la construcción de modelos estadísticos, de modo que permitan la separación, con el menor error posible, de las diferentes tipologías de este fenómeno. Como resultado de esta aplicación fue posible predecir 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, género, estado civil, recursos económicos, tipo de colegio, orientación profesional recibida y nivel de estudios de los padres entre otros) con un error mínimo del 29% para un total de 18 variables. Ésta última tarea, conocida como predicción de clases permitió identificar qué características de los estudiantes nuevos se traducirían en un alto riesgo de abandono (abandonos potenciales).

**Descriptor o Palabras Clave:** Abandono, Clasificación de Patrones y Riesgo.

## Introducción

Podría decirse que uno de los principales retos que 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- el aumento en la demanda, la política ampliación de cobertura de la educación superior, el número de estudiantes que logran culminar sus estudios en el tiempo previsto, los altos niveles de abandono y las bajas tasas de graduación, imponen restricciones importantes al impacto de las políticas de permanencia.

Aunque las causas y consecuencias del abandono estudiantil universitario han sido ampliamente estudiadas para explicar por qué un estudiante toma la decisión de abandonar sus estudios a lo largo de su ciclo académico como resultado de la interacción de diferentes determinantes (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), 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]), no se dispone de aproximaciones empíricas que permitan su correcta clasificación y predicción, lo que dificulta la tarea de implementar políticas de permanencia.

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 su aplicación por el Proyecto Alfa-GUÍA, ha sido posible aplicar la teoría de clasificación (también conocida como clasificación de patrones) en la construcción de modelos estadísticos que permiten 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. Derivado de estos modelos es posible usar las 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. 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 cuatro secciones. En la segunda, se describe brevemente la metodología en términos del tipo de datos, y los modelos de clasificación de patrones. En la tercera se presentan los resultados para las regiones en las que tiene sedes

Universidad de Antioquia y en la cuarta las conclusiones.

## 2. Metodología

### 2.1. 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). Toma información de 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.

- **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.

En total se encuestaron 409 de las regiones donde tiene sede la Universidad de Antioquia. De las encuestas aplicadas sólo se obtuvieron datos confiables para 371, lo que representa una pérdida del 9.29% de las observaciones.

**Tabla 1.** Total de estudiantes por tipologías de abandono

Sede	Tipología					Subtotal <sup>a</sup>	Total
	Ac	CP	CIES	CMN	Ab		
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
Urabá	11	1	4	1	4	10	21
Suroeste	29	3	6	0	14	23	52
Sonsón	2	0	0	0	3	3	5
Total	166	29	66	20	90	371	371

<sup>a</sup> Suma de CP+CIES+CMN+AB

## 2.2. 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. Así, se observa un conjunto de  $n$  parejas  $\{(x_i; y_i); i = 1, \dots, n\}$  independientes e idénticamente distribuidas (i.i.d.) acorde con una función de distribución de probabilidad desconocida  $\mathbb{P}(x, y)$ , donde  $x_i \in \mathbb{R}^+$  y 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,  $K = 2$ .

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),  $y \in \{1, 2, \dots, K\}$ . Así, en la encuesta se indagó por las cinco (5) tipologías descritas.

En consecuencia, dado el conjunto de entrenamiento  $\{(x_i, y_i); i = 1, \dots, n\}$ , el objetivo del pro-

blema de clasificación consiste en aprender la regla de decisión óptima.

$$\phi(x) = \operatorname{argmax}_{k=1, \dots, K} f_k(x)$$

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

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.

Además, para la correcta aplicación de la metodología de clasificación se identificaron las variables predictoras relevantes, esto es, cuáles de las  $p$  variables en el vector  $x$  son importantes para construir las funciones de clasificación. La selección de variables es necesaria cuando el vector  $x$  es de alta dimensión (i.e. el número de variables  $p$  es grande con respecto al número de observaciones  $n$ ) y cuando el problema de clasificación es de múltiples clases

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  $\beta$  asociado al vector de variables predictoras  $x$ , con el fin de identificar cuáles coeficientes  $\beta_j$  asociados a las variables  $x_j$ ,  $j = 1, \dots, p$  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.  $J(\beta) = \{j \in \{1, \dots, p\} : \beta_j \neq 0\}$ , 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  $x \in \mathbb{R}^+$  en el sentido que  $\hat{\beta}(\lambda) = 0$  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  $\lambda$  mayor restricción sobre el número de variables a incluir en el modelo. En la práctica, el parámetro  $\lambda$  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]).

Además, el 70% de las observaciones fue seleccionada aleatoriamente (sin reemplazo) para entrenar las reglas de clasificación, y el 30% restantes fue utilizado 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 la Tabla 2.

**Tabla 2:** Distribución de datos por clases y por regularización

Clases	Datos			
	Entrenamiento	Prueba	Total	
K = 2	Activo	118	56	174
	Abandono	159	64	223
	Total	277	120	397
K = 5	Activo	119	55	174
	Cambio Programa	23	9	32
	Cambio IES	53	19	72
	Cambio Nivel	13	8	21
	Abandono	69	29	98
Total	277	120	397	

### 3. Metodología

#### 3.1. Modelos Logísticos

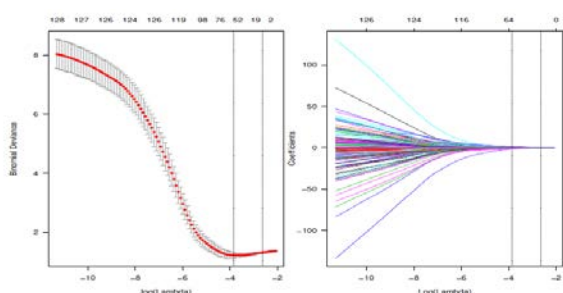
Según los resultados obtenidos se cuenta con un total máximo de 59 variables para estudiar el abandono separado en dos clases y 28 variables para estudiar las cinco (5) tipologías de abandono con los errores de predicción que se muestran en la tabla 3.

**Tabla 3.** Errores de predicción

Parámetro de Regularización $\lambda$	$\log(\lambda)$	Error	
		$\lambda_{\min}$	$\lambda_{1se}$
K = 2	$\lambda_{\min} = 0.021$	-3.858	0.425
	$\lambda_{1se} = 0.071$	-2.649	0.4
K = 5	$\lambda_{\min} = 0.0345$	-3.365	0.5
	$\lambda_{1se} = 0.0662$	-2.714	0.542

Al comparar los resultados obtenidos por  $\lambda_{\min}$  y  $\lambda_{1se}$  se encontraron algunos factores que podrían denominarse protectores para la permanencia y que podrían clasificarse en previos a la entrada a la IES y aquellos que se adquieren con la interacción del ambiente académico e institucional. En los factores que podrían considerarse posteriores a la entrada a la IES y que podrían disminuir el riesgo de abandono se identificaron: estar matriculado carreras presenciales en comparación con aquellos que cursan pregrados a distancia, el recibir apoyo económico en forma de becas o subsidios o trabajo en la institución, el 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 conyugue 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.



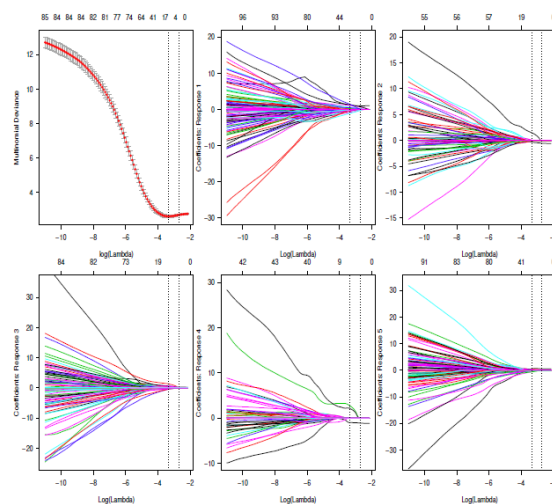
**Figura 1:** Gráfica izquierda: validación cruzada de 10-iteraciones. Gráfica derecha: trayectorias de los coeficientes (con penalización  $l_1$ ) estimados. La línea vertical izquierda corresponde al mínimo error, mientras que la línea vertical derecha corresponde al mayor valor de  $\lambda$  tal que el error esté dentro de un error estándar del valor mínimo.

De otro lado se tienen aquellos factores de parecen generar un mayor riesgo de abandono tales como: no recibir ningún tipo de ayuda económica, haber vivido alguna experiencia traumática dentro de la institución y vivir con el conyugue.

En cuanto al modelo multinomial (cinco clases) se obtuvo 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.

Es importante anotar inicialmente que el error obtenido del análisis de clasificación fue superior al 0,40, en este caso el modelo binomial, independientemente del  $\lambda$ , genera mayor riesgo de permanencia para aquellos estudiantes que viven con familiares, reciben apoyo económico en forma de becas, subsidios o trabajo institucional. 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.

En cuanto al modelo multinomial, 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.



**Figura 2:** 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 línea vertical derecha corresponde al mayor valor de  $\lambda$  tal que el error esté dentro de un error estándar del mínimo.

#### 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 y explicarían 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. Un análisis factorial confirmatorio o un análisis por componentes principales permitiría construir un índice de abandono donde se le asigne a cada variable el peso dentro del fenómeno e identificar el riesgo de cada estudiante en una escala de 0 a 1, esto permitiría identificar los factores de mayor peso en la clasificación como estudiantes que abandonan o como estudiante que permanece.

## Referencias

- Tinto, V. (1989). Definir la deserción: una cuestión de perspectiva, *Revista de Educación Superior*, 71, México.
- Tinto, V. (1975). Dropout From Higher Education: A Theoretical Synthesis of Recent Research, *Review of Education Research*, 45, 89-125.
- Bean, J. P. (1980). Student Attrition, Intentions and Confidence. *Research in Higher Education*, 17, 291-320.
- Cabrera, A., Nora, A y Castañeda, M., (1993). College Persistence: Structural Equations Modelling Test of Integrated Model of Student Retention, *Journal of Higher Education*, 64(2).123-320.
- Spady, W. (1970). Dropout from Higher Education: An Interdisciplinary Review and Synthesis, *Interchange* 1, 64-85.
- Porto, A., y Di gresia, L., (2001). Rendimiento de estudiantes universitarios y sus determinantes” *Asociación Argentina de Economía Política*, Noviembre.
- Cornwell, C. (2002). The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia’s HOPE Scholarship. University of Georgia, Department of Economics
- Cameron, S y Taber, C. (2001). Estimation of Education Borrowing Constraint using Returns Schooling. NBER working paper, No.W7761.
- DesJardins, S., Ahlburg, D., and McCall, B., (2001). Simulating the Longitudinal Effects of Ghanges in Financial Aid on Student Departure from College. *Journal of Human Resources*, 37, 653-679.
- DesJardins, S., Ahlburg, D., and McCall, B., (2002). A Temporal Investigation of Factors Related to Timely Degree Completion. *The Journal of Higher Education*, 73, 555-581.
- Alemany, R., (1990). Modelación de la duración de estudios universitarios: una aplicación a la universidad de Barcelona. Tesis doctoral, Universidad de Barcelona.
- Cameron, S. and Heckman, J., (1998). Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. *The Journal of Political Economy*, 106, 262-333.
- Booth, A. and Satchell, S., (1995). 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.
- Häkkinen, L. and Uusitalo, R., (2003). The Effect of a Student Aid Reform on Graduation: A Duration Analysis. Working Paper Series No. 8, Department of Economics, Uppsala University.
- Willett, J.B. and Singer, J.D., (1991). From Whether to When: New Methods for Studying Student Dropout and Teacher Attrition. *Review of Educational Research*, 61, 407-450.
- Giovagnoli, P. (2002). Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración,” *Documento de Trabajo 37*, Universidad Nacional de la Plata
- Hastie, T; Tibshirani, R and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 5th edition.
- Tinto, V. (1982). Limits of theory and practice of student attrition. *Journal of Higher Education*, 53 (6), 687-700
- Singer, J.D. and Willett, J.B., (1993). It’s About Time: Using Discrete-Time Survival Analysis To Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18, 155-195
- Psacharopoulos, G., 1985. “Returns to education: A Further Update and Implications,” *The Journal of Human Resources*, 20, 583-604.
- Psacharopoulos, G., and Patrinos, H., 2002. “Returns to Investment in Education: A Further Update,” *Policy Research Working Paper*, 2881, World Bank