Revista d'Innovació Docent Universitària
Núm. 10 (2018), pp. 13-24
http://revistes.ub.edu/index.php/RIDU

Factores socio-académicos asociados a la tasa de abandono en el Grado de Estadística

Socio-academic factors associated to the drop-out rate in the Degree of Statistics

Manuela Alcañiz Zanón*, Lluís Bermúdez Morata, Sandra García Martínez***, Jordi López Tamayo***

*Departament d'Econometria, Estadística i Economia Aplicada
Universitat de Barcelona
Diagonal, 690, 08034 Barcelona.
**Departament de Matemàtica Econòmica, Financera i Actuarial
Universitat de Barcelona
Diagonal, 690, 08034 Barcelona.
***Departament d'Epidemiologia i Estadística
Salut de la Dona Dexeus.
malcaniz@ub.edu lbermudez@ub.edu sandra11441@hotmail.com jlt_lopez@ub.edu

Resumen

La tasa de abandono es uno de los indicadores más utilizados en la valoración de los estudios de Grado universitarios. Analizar esta variable y los factores que influyen en ella puede aportar información de interés a tener en cuenta tanto por parte de la Dirección y profesorado del Grado, como por los propios estudiantes.

Se parte de la información facilitada por el Servicio de Planificación Académico-docente de la Universidad de Barcelona de los estudiantes que han iniciado el Grado de Estadística, impartido conjuntamente por la Universidad de Barcelona (UB) y la Universidad Politécnica de Cataluña (UPC), desde 2009 a la actualidad. Se analiza qué factores socio-académicos de los nuevos estudiantes del Grado pueden influir en la probabilidad de abandono de éste. La metodología utilizada para ello ha sido un modelo de regresión logística (lógit).

La vía y la nota de acceso al Grado, y el nivel de estudios de los padres, influyen significativamente en la tasa de abandono del Grado de Estadística. Concretamente: (i) los estudiantes que acceden al Grado a través de los ciclos formativos de grado superior presentan una tasa de abandono mayor que aquellos que lo hacen desde el bachillerato; (ii) una nota de acceso elevada disminuye la tasa de abandono y, por último, (iii) los estudiantes cuyos padres no tienen formación universitaria tienden a abandonar menos sus estudios de Grado.

Palabras clave: Tasa de abandono, Estudios universitarios**,** Modelo lógit, Matriz de confusión, Curva ROC.

Abstract

Drop-out rate is one of the most common indicators used by assessment programmes for university degrees. Analyse this rate and its key influencing factors may be reveal information of interest to take into consideration by the Direction of the Degree and by the students as well.

Using the information, provided by the Servei de Planificació Acadèmic-docent of the University of Barcelona, on the students that have initiated the Degree of Statistics from 2009 to now, this study analyses which socio-academic factors can influence in the probability of dropping out this Degree. The methodology used for this purpose has been a logistical regression model (logit).

University access route, university entrance grade, and educational level of the parents, influence significantly in the drop-out rate of the Degree of Statistics. Specifically: (i) students accessing to the Degree through upper grade

DOI: http://dx.doi.org/10.1344/RIDU2018.10.2

educational cycle present a larger drop-out rate than those that do it from high school degree; (ii) a larger university entrance grade diminishes the drop-out rate and, finally, (iii) students whose parents do not have university degrees tend to have a smaller drop-out rate.

Key words: Drop-out rate, University degrees, Logit model, Confusion matrix, ROC curve.

1. Introducción

Actualmente nos encontramos en una sociedad en que la formación académica es fundamental para acceder al mundo laboral (Brennan, 2008). Las altas tasas de paro registradas en España hacen que las empresas sean cada vez más exigentes con el perfil de los candidatos a formar parte de sus plantillas; concretamente, cada vez se piden estudios de mayor nivel para ocupar puestos de trabajo no siempre acordes con ellos. Se produce así con frecuencia una sobreeducación que puede suponer una trampa para la persona que la sufre, y que tiene consecuencias sociales muy negativas (Rubb, 2003).

Al margen de este fenómeno, la importancia de disponer de una o varias titulaciones universitarias es indiscutible en el momento de encontrar trabajos de calidad (Lassibille *et al*., 2001). Todos los estudios, ya sean secundarios, grados, masters, etc., tienen una duración predeterminada. No obstante, hay muchos estudiantes que no terminan en este tiempo previsto, ya sea porque tardan más o menos en finalizarlos, o simplemente porque no los terminan nunca a causa del abandono. Este fenómeno ha sido estudiado ampliamente (Álvarez Pérez *et al.*, 2006; Bernardo *et al*., 2016), dado que da lugar a muchas consecuencias negativas, por ejemplo, pagar matrículas durante más años, sufrir el incremento del precio del crédito al repetir asignaturas, demorar el momento de finalización de la etapa académica y la incorporación al mercado laboral, etc. Desde el punto de vista personal del estudiante, el abandono de unos estudios y la indefinición que supone elegir otros o renunciar a seguir estudiando, tienen también costes en términos de autovaloración y confianza en sus capacidades (Park *et al*., 2007).

Sin embargo, centrándonos en el abandono de los estudios universitarios, su incidencia y las características que lo explican son peculiares para cada titulación. En este contexto, el presente artículo aborda el análisis de la tasa de abandono por parte de los estudiantes que realizan el Grado de Estadística impartido de manera conjunta por la UB y la UPC. Mediante un modelo de regresión logística, se identificarán, entre un conjunto de variables socio-académicas, qué factores influyen significativamente en la probabilidad de abandono de los estudios de Grado.

A partir de los resultados, se discutirán las conclusiones que pueden derivarse de estos. En particular, se valorarán las conclusiones para cada uno de los agentes implicados en la titulación. En primer lugar, se proporcionará a la Dirección del Grado una valiosa información sobre los motivos que abocan al abandono a determinado perfil de estudiante. Ello permitirá diseñar estrategias a tiempo para evitar la pérdida de estudiantes. En segundo lugar, al profesorado del Grado se le proporcionará un mayor conocimiento sobre el perfil y debilidades de los estudiantes. Ello puede ayudar a ajustar contenidos, a suplir carencias previas de conocimiento, a ejercer una buena tutorización, etc. Finalmente, el propio estudiante puede ser más consciente de sus puntos débiles. El conocimiento a tiempo de ellos puede ayudarle a planificar con más realismo sus horas de estudio, su asistencia a clase, etc., de modo que su rendimiento sea suficiente para seguir adelante con sus estudios aunque sus condiciones de acceso a la titulación no fueran las mejores.

A modo de presentación, en la siguiente sección se comenta la obtención, tratamiento y descripción de los datos utilizados en este trabajo. A continuación, en la sección 3 se presenta la metodología estadística utilizada para el análisis de los datos disponibles. En la sección 4, se presentan los resultados obtenidos y, finalmente, en la sección 5, se detallan las conclusiones alcanzadas.

2. Datos

Para llevar a cabo los objetivos planteados, ha sido necesaria la obtención de una base de datos con la información de los estudiantes matriculados en el Grado de Estadística. En este caso, los datos han sido aportados por el Servicio de Planificación Académico-docente de la Universidad de Barcelona.

La base de datos contiene información sobre los individuos matriculados en el Grado de Estadística, desde el año 2009 (inicio del grado) hasta el año 2013, y del seguimiento que de estos estudiantes se ha realizado hasta el curso 2016-2017. Se dispone de información de un total de 251 estudiantes. De la información disponible en la base de datos se han seleccionado las variables socio-académicas relacionadas con la tasa de abandono de los estudios.

En primer lugar, se ha definido la variable de interés, denominada *abandono*. Esta variable dicotómica es un indicador que toma valores 0 o 1, 1 en caso de que el individuo lleve 2 o más cursos sin matricularse y 0 en caso contrario. En el caso del Grado de Estadística interuniversitario UB-UPC, se aplica por convenio la Normativa de Permanencia de la UB para estudiantes que cursan grados o masters universitarios[1]. Esta normativa recoge en el artículo 6.1. que se considera que un alumno ha abandonado un grado cuando han transcurrido dos años consecutivos sin que se haya matriculado. Podemos avanzar que, según esta definición, 124 estudiantes han abandonado el Grado, lo que representa un 49,40% del total.

En segundo lugar, las variables tenidas en cuenta para analizar su incidencia en la tasa de abandono han sido divididas en dos grupos. Las variables *sexo, edad, domicilio, estudios padres* y *trabaja* forman parte de los factores sociológicos. Por otra parte, las variables *vía de acceso, nota de acceso, orden* y *curso inicio*, forman parte de los factores académicos. Para cada individuo la información relativa a estas variables es la información recogida en el momento de su primera matriculación en el Grado. En la Tabla 1 se muestra la descripción de las variables analizadas.

Respecto a las variables clasificadas como sociológicas, la variable *sexo* es una variable dicotómica formada por dos categorías, Hombre y Mujer; la variable *edad* ha sido categorizada en dos grupos atendiendo a la edad fijada para el acceso a la Universidad de los mayores de 25 años; la variable *domicilio* ha sido categorizada en dos grupos, diferenciando los estudiantes domiciliados en los municipios pertenecientes a la Área Metropolitana de Barcelona (AMB) del resto; en la variable *estudios padres*, se ha diferenciado entre aquellos padres con al menos uno de ellos con estudios universitarios del resto; por último, la variable *trabaja* es una variable dicotómica que indica si el estudiante trabaja o no.

Respecto a las variables académicas, la variable *vía de acceso* divide a los alumnos según si estos han accedido a los estudios a través de la selectividad (PAU), de los ciclos formativos de grado superior (CFGS) o a través de una Licenciatura previa; la variable *nota de acceso* ha sido categorizada en tres grupos, los estudiantes con nota de acceso inferior o igual a 7, con nota entre 7 y 9 y con nota superior a 9; la variable *orden* hace referencia al orden de preferencia escogido por el estudiante, diferenciando los estudiantes que escogieron el Grado de Estadística como primera opción de aquellos que no; por último, la variable *curso inicio* indica el año de la primera matrícula en el Grado de Estadística.

A partir de la Tabla 1 se pueden extraer conclusiones sobre el perfil del estudiante del Grado de Estadística. Casi un tercio de estudiantes son hombres. Solo un 8,76% de los alumnos que empiezan el Grado tienen más de 25 años. Un 66,93% tienen la residencia familiar en el área metropolitana de Barcelona, próxima a la Facultad donde se imparten las clases. Un 74,10% de los estudiantes no trabajan cuando inician estos estudios y, por último, un 60,16% tienen padres sin ningún tipo de estudios universitarios.

[1] Disponible en: http://www.ub.edu/acad/noracad/permanencia.pdf.

En cuanto a las características académicas de estos estudiantes, se observa que la gran mayoría de los estudiantes acceden al Grado de Estadística a través de las pruebas de acceso a la universidad (PAU) y escogiendo estos estudios como primera opción. Un 56,97% entran con una nota menor a 7, nota de acceso no muy alta teniendo en cuenta que puede haber como máximo una nota de 14.

Finalmente, en la última columna, la Tabla 1 muestra la tasa de abandono para cada una de las categorías analizadas. En todos los casos se puede observar como la tasa de abandono es bastante elevada.

En este estudio, la explotación de datos se ha realizado con el software estadístico R. R es un entorno y lenguaje de programación con un enfoque al análisis estadístico de datos.

Tipo de variable	Variable	Categoría	% Alumnos	% Abandono
Factores sociológicos	*Sexo*	Mujer	37,45	50,00
		Hombre	62,55	49,04
	Edad	< 25 años	91,24	47,59
		≥ 25 años	8,76	68,18
	Domicilio	AMB	66,93	50,59
		Resto	33,07	46,99
	Estudios padres	Universitarios	39,84	56,00
		No univ.	60,16	45,03
	Trabaja	Trabaja	25,89	55,38
		No trabaja	74,10	47,31
Factores académicos	*Vía de acceso*	PAU	88,45	46,85
		CFGS	8,37	71,43
		Licenciados	3,19	62,50
	Nota de acceso	≤ 7	56,97	58,04
		(7,9]	29,48	45,94
		>9	13,55	20,59
	Orden	1r orden	89,24	47,77
		+ de 1r orden	10,76	62,96
	Curso inicio	2009	18,32	56,52
		2010	22,71	66,66
		2011	17,53	54,54
		2012	19,92	32,00
		2013	21,51	37,04

Tabla 1: Descripción y resumen estadístico de las variables explicativas

3. Metodología

En este apartado de metodología, se detallará de manera más específica cuáles han sido los métodos utilizados para tratar y analizar los datos disponibles. La principal técnica estadística utilizada ha sido el análisis logístico. En primer lugar, se ha estimado una regresión logística. Posteriormente, para la validación de dicha regresión logística, se ha obtenido la matriz de confusión y la curva ROC.

3.1. Modelo de regresión logística

La regresión logística es una de las técnicas más conocidas y utilizadas para modelizar una variable respuesta categórica en función de unas variables independientes continuas o categóricas (Peng *et al*., 2002). El objetivo de este modelo se puede dividir en tres fases: i) determinar la existencia o ausencia de relación entre una o más variables independientes y una variable dicotómica, ii) medición de dicha relación en caso de que exista, y iii) estimar o predecir la probabilidad de que se produzca el suceso en función de los valores que adoptan las variables independientes.

La variable dependiente *Y* puede tomar dos valores, codificados usualmente como 1 para la categoría de interés y 0 para la categoría contraria. De esta manera, la distribución de *Y* es una Bernoulli con la siguiente esperanza:

$$E[Y]=P[Y=1]=p\ (0<p<1).$$

Se puede tener una o más variables *X*, posibles predictores de la variable *Y*, entonces la distribución condicional de *Y* sobre estos valores, también siguen una distribución Bernoulli de forma que su esperanza y varianza condicionadas de *Y* sobre *X* (suponiendo que solo tenemos un predictor) es:

$$E[Y|X=x]=P[Y=1|X=x]=p(x),$$
$$Var[Y|X=x]=p(x)\cdot(1-p(x)).$$

Aplicando un modelo gaussiano de regresión lineal se obtendría:

$$E[Y|X=x]=p(x)=\alpha+\beta x+\varepsilon.$$

No obstante, este modelo gaussiano presenta algunos problemas como la falta de normalidad, heterocedasticidad, limitaciones de los valores de *p(x)*, etc. Por este motivo, se utilizan modelos de regresión alternativos con la siguiente estructura:

$$Y=F(\alpha+\beta x)+\varepsilon$$

dónde *F* es una función monótona creciente. En este estudio, la función seleccionada ha sido la transformación lógit, que da lugar al modelo de regresión logística simple:

$$\text{logit}(p(x))=\ln\left(\frac{p(x)}{1-p(x)}\right)=\alpha+\beta x.$$

Con dicha transformación, *p(x)* está acotada entre el 0 y el 1, como deber ser puesto que *p(x)* es una probabilidad.

En este estudio se realiza un modelo de regresión logística múltiple, al considerarse *R* variables explicativas $X_1,...,X_R$. De esta manera, la fórmula que interesa modelizar es la siguiente:

$$E[Y|X_1=x_1,...,X_R=x_R]=P[Y=1|X_1=x_1,...,X_R=x_R]=p(x_1,\ldots,x_R).$$

Con la transformación lógit se obtiene:

$$\ln\left(\frac{p(x)}{1-p(x)}\right)=\alpha+\sum_{r=1}^{R}\beta_r x_r.$$

Es decir, la regresión logística predice la probabilidad de ocurrencia de un suceso ajustando los datos a una función lógit.

En concreto, para este trabajo, la variable dependiente *Y* es la variable *abandono*, con la que sabemos con exactitud si el individuo ha abandonado o no los estudios de Estadística. Las posibles variables explicativas *X* son el conjunto de variables socio-académicas descritas en la Tabla 1.

3.2. Estimación y selección del modelo

Tal y como se ha comentado en el subapartado anterior, se pueden tener *R* variables explicativas o independientes para estudiar el abandono del Grado por parte de los estudiantes de Estadística. El modelo a utilizar presentaría la siguiente estructura:

$$\text{ABANDONO} = f(X_1, X_2, \ldots, X_R).$$

El modelo de regresión logística a estimar sería el siguiente:

$$\ln [P(\text{ABANDONO}=1) / P(\text{ABANDONO}=0)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_R X_R .$$

Una vez ajustado el modelo de regresión logística a los datos, se podrá observar qué variables inciden significativamente en la probabilidad de abandonar el Grado de Estadística y qué variables son independientes.

Para determinar el modelo que mejor se ajuste a los datos, el criterio de selección utilizado será tener en cuenta el nivel de significación estadístico de las variables y el criterio de Akaike (AIC; Bozdogan, 1987). El AIC es una medida de ajuste que penaliza el modelo por el número de coeficientes del mismo. De esta manera, interesa un modelo donde este criterio de información sea el menor posible.

En un modelo de regresión logística, una vez seleccionadas las variables significativas, la expresión que refleja la probabilidad de abandono se define de la siguiente manera:

$$P(\text{ABANDONO} = 1) = \frac{1}{1 + \exp\left(-\left(\alpha + \sum_{r=1}^{R} \beta_r x_r \right) \right)}.$$

3.3. Validación del modelo

Una vez realizada la estimación del modelo que mejor se ajusta a los datos disponibles, se procede a la validación del mismo. La validación indica la precisión con la cual un modelo se aproxima a los datos observados. Para evaluar la idoneidad del modelo se tendrán en cuenta la matriz de confusión y la curva ROC, tal y como se ha comentado al inicio del apartado.

La curva ROC es un gráfico donde se pueden observar todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados.

Así, los conceptos de sensibilidad y especificidad son esenciales para entender la curva ROC. Este método se utiliza de manera usual para los ensayos clínicos. La sensibilidad es la razón de verdaderos positivos, es decir, en el caso de una prueba de diagnosis, la sensibilidad es la probabilidad de obtener un resultado positivo cuando el individuo tiene la enfermedad. Mide la capacidad para detectar la enfermedad cuando está presente. Por otro lado, la especificidad es la razón de verdaderos negativos, es decir, la probabilidad de obtener un resultado negativo cuando el individuo no tiene la enfermedad y por tanto mide la capacidad para descartar la enfermedad cuando esta no está presente.

Para mostrar más fácilmente estos conceptos se tiene en cuenta la matriz de confusión. La tabla de confusión o la matriz de confusión contiene información sobre las predicciones realizadas por un método o sistema de clasificación. Entonces, como puede verse en la Tabla 2, cada columna de la matriz

representa el número de predicciones de cada clase, mientras que cada fila representa las instancias en la clase real.

		Predicción	
		Éxito	**Fracaso**
Real	**Éxito**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Fracaso	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Tabla 2: Matriz de confusión para dos clases

Con esta tabla se puede calcular el porcentaje de individuos que el modelo clasifica correctamente. Interesa un porcentaje elevado ya que este hecho significará que el modelo es adecuado. A partir de la información contenida en la matriz de confusión, la sensibilidad y la especificidad tienen las fórmulas siguientes:

- Sensibilidad: verdaderos positivos / total éxito = $\frac{VP}{VP+FN}$.
- Especificidad: verdaderos negativos / total fracaso = $\frac{VN}{VN+FP}$.

A partir de estos conceptos, como puede verse en la Figura 1, en la curva ROC la sensibilidad se sitúa en el eje de ordenadas del gráfico, mientras que en el eje de abscisas se sitúa 1-especificidad.

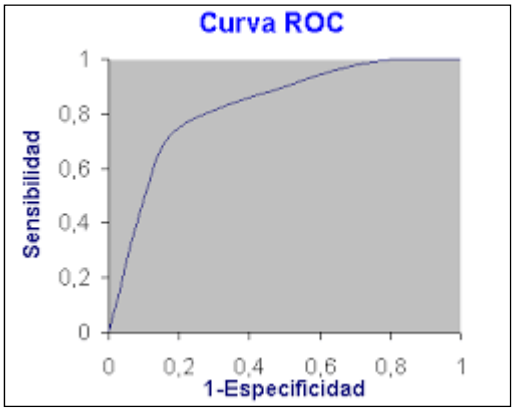

Figura 1: Ejemplo de una curva ROC

Teniendo en cuenta el gráfico, la prueba diagnóstica ideal debería tener una sensibilidad y una especificidad lo más próximas posibles a 100%, es decir, cuanto más próxima esté la curva a la esquina izquierda superior, más alta es la exactitud global de la prueba.

Se considera que un modelo es mejor que otro modelo siempre y cuando el área bajo la curva sea mayor. Esta área bajo la curva (AUC) puede estar entre 0,5 y 1 donde 1 representa un valor diagnóstico perfecto mientras que 0,5 es una prueba sin capacidad discriminatoria diagnóstica.

4. Resultados

Como se ha comentado en el apartado anterior, la regresión logística es adecuada cuando se quiere relacionar una variable dependiente dicotómica con una o más variables independientes. En el estudio

que nos ocupa, la variable dependiente será el indicador *abandono* que como ya sabemos toma valores 1 en caso de abandono, 0 en caso contrario. En la base de datos analizada, de un total de 251 individuos, 124 estudiantes están clasificados como estudiantes que han abandonado el Grado.

Las variables independientes o explicativas que se han tenido en cuenta en el modelo de regresión logística son las variables descritas en la Tabla 1. Por un lado, las características sociológicas del estudiante: el sexo, la edad, el domicilio, el nivel de estudios de los padres y su situación laboral. Por otro lado, relacionado con los aspectos académicos del estudiante, la vía y la nota de acceso al Grado, el orden de preferencia del Grado en las opciones escogidas por el estudiante y el curso de inicio de los estudios de Grado.

El objetivo de este estudio es determinar qué características influyen significativamente en la tasa de abandono y analizar de qué manera lo hacen. La presentación de los resultados se puede dividir en dos apartados: estimación y validación del modelo.

4.1. Estimación del modelo

En esta primera parte se quiere determinar el modelo que mejor se ajuste a los datos. Para llegar a este objetivo, se tendrá en cuenta el nivel de significación estadístico de las variables y el criterio de Akaike (AIC).

Una vez estimado el modelo de regresión logística que incluye todas las variables explicativas mencionadas anteriormente, se concluye que las variables *sexo*, *edad*, *domicilio*, *trabaja* y *orden* no son significativas estadísticamente.

Así pues, el modelo final, con mejor AIC, incluye las variables explicativas siguientes: *estudios padres*, *vía de acceso*, *nota de acceso*, y *curso inicio*. Todas estas variables presentan unos coeficientes estadísticamente significativos en alguna de sus categorías. Por tanto, se puede concluir que son variables relevantes en el análisis de la tasa de abandono. En la Tabla 3 se presentan los resultados de dicho modelo.

Variable	**Coeficiente**	**Odds ratio**	**Error estándar**	**p-valor**
Constante	0,7863	2,1953	0,3705	0,0338*
Estudios padres (no univ.)	-0,7347	0,4797	0,2941	0,0125*
Vía de acceso (CFGS)	1,3963	4,0403	0,5342	0,0089**
Vía de acceso (Licenc.)	0,7108	2,0355	0,7796	0,3619
Nota de acceso ((7,9])	-0,9716	0,3785	0,3408	0,0044**
Nota de acceso (>9)	-2,1514	0,1163	0,5431	0,0000**
Curso inicio (2010)	1,2356	3,4404	0,4746	0,0092**
Curso inicio (2011)	0,0631	1,0652	0,4419	0,8863
Curso inicio (2012)	-0,7298	0,4820	0,4519	0,1064
Curso inicio (2013)	-0,5325	0,5871	0,4447	0,2311
AIC	316,82			

* Significación al 5%, ** significación al 1%.

Tabla 3: Resultados del modelo final de regresión logística

Los valores de los odds ratio (exponencial del coeficiente) son los que ofrecen información sobre la influencia de la variable sobre el abandono. Para este modelo, el individuo de referencia es un estudiante matriculado por primera vez en el curso 2009-2010, con algún padre con estudios universitarios, con

acceso al Grado a través de las pruebas de accesos a la universidad y una nota inferior a 7. Respecto a este individuo de referencia se realizan las comparaciones.

La variable *estudios padres* es significativa, el p-valor asociado a la segunda categoría de esta variable, padres con estudios no universitarios, es de 0,0125 lo que indica significación estadística al 5%. Atendiendo al valor del odds ratio, la probabilidad de abandonar de los estudiantes con padres sin estudios universitarios se reduce en un 52,03% respecto de los estudiantes cuyos padres poseen estudios universitarios.

De todas las categorías de la variable vía de acceso, el coeficiente de la categoría relativa a los estudiantes que han accedido al Grado de Estadística a través de los ciclos formativos de grados superiores posee significación estadística con un p-valor asociado de 0,0089. Los estudiantes que han accedido a este grado universitario por CFGS, tienen una probabilidad de abandonar 4 veces mayor que los estudiantes que han accedido a través de las pruebas de acceso a la universidad (PAU).

Las categorías de la variable nota de acceso presentan también unos p-valores asociados menores que el nivel de significación del 5%. Estas categorías obtienen unos coeficientes negativos que implican odds ratios inferiores a 1 y, por tanto, menor probabilidad de abandono para notas de acceso más altas que las de la categoría de referencia (nota entre 5 y 7). De manera específica, si el estudiante accede a la universidad con una nota entre 7 y 9, la probabilidad de abandonar se reduce en un 62,16%. Por otro lado, si el estudiante tiene una nota de acceso superior a 9, la probabilidad de este indicador de abandono se reduce en un 88,37%. Esta interpretación es lógica ya que se espera que cuanta más nota tenga un estudiante, menos probabilidad de abandonar tendrá.

La categoría Curso inicio 2010 (estudiantes que han iniciado los estudios el año 2010) posee significación estadística con un p-valor asociado de 0,0092. Los estudiantes que iniciaron sus estudios en este curso, tienen una probabilidad de abandonar la carrera 3,44 veces más que los estudiantes que han iniciado el curso el año 2009. El resto de cursos no presentan diferencia significativa respecto del año 2009.

Una vez interpretados todos los coeficientes de las variables, se observa el criterio de Akaike el cual toma un valor de 316,82. Anteriormente a este modelo descrito, se han realizado pruebas con modelos auxiliares cuyos criterios de Akaike eran superiores al criterio de este mejor modelo estimado, por lo que fueron descartados.

4.2. Validación del modelo

Una medida de la bondad de ajuste es un estadístico resumen que indica la precisión con la cual un modelo se aproxima a los datos observados. Para evaluar la idoneidad de este modelo, estimado mediante la regresión logística, se tendrán en cuenta las dos medidas estadísticas ya mencionadas: la matriz de confusión y la curva ROC.

El ajuste del modelo analiza en qué medida los factores introducidos en la especificación del modelo son capaces de explicar el comportamiento de la variable dependiente. Este ajuste vendrá explicado por el porcentaje de casos correctamente clasificados por el modelo estimado dentro de su categoría correspondiente. La Tabla 4 muestra la matriz de confusión del modelo final estimado.

		Predicción		
		No abandonan	**Abandonan**	**Total**
Real	**No abandonan**	73	54	127
	Abandonan	26	98	124
	Total	99	152	251

Tabla 4: Matriz de confusión del modelo final

Una vez estimado el modelo, se realiza la clasificación de los individuos. Teniendo en cuenta el ajuste del modelo para cada individuo, se crean las categorías 0 y 1 las cuales hacen referencia al indicador *abandono*. Si el ajuste del modelo para un individuo es inferior a 0,5 entonces se le asignará un 0 y en caso de ser mayor a 0,5, se le asignará un 1. Un total de 171 estudiantes se han clasificado correctamente en sus categorías respectivas y por tanto el porcentaje de aciertos es del 68,13%.

El análisis por categorías muestra la existencia de un 57,48% de estudiantes que no han abandonado la carrera de estadística bien clasificados por el modelo (sensibilidad); y un 79,03% de estudiantes que sí que han abandonado clasificados también correctamente (especificidad). En general, el modelo sobrestima ligeramente el número de estudiantes que abandonan los estudios.

El error que se puede encontrar son aquellos estudiantes que han sido clasificados incorrectamente en categorías que no les correspondía. En este caso, 54 estudiantes que no abandonaron sus estudios fueron clasificados por el modelo como estudiantes que abandonan, mientras que 26 estudiantes que abandonaron realmente sus estudios fueron clasificados por el modelo como estudiantes que no abandonan.

En la Figura 2 se muestra la curva ROC para el modelo final estimado. La interpretación de este indicador se realiza observando el área por debajo de la curva. Interesa que la curva esté lo más arriba de la izquierda posible. Si la curva estuviera sobre la línea diagonal indicaría que el ajuste no es bueno y que por tanto el modelo no ajusta bien los datos.

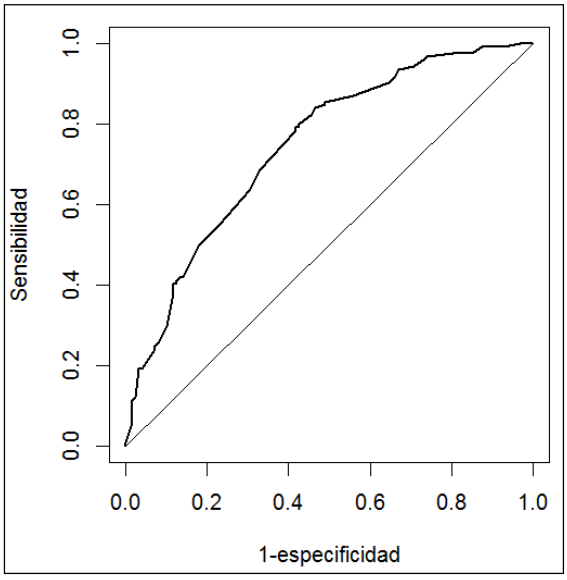

Figura 2: Curva de ROC del modelo final

La Figura 2 muestra como la curva está por encima de la línea diagonal. De manera específica, el área bajo la curva es de 0,75, valor que puede considerarse como aceptable. Así pues, el modelo final, sin ser excelente, se ajusta aceptablemente a los datos disponibles.

5. Conclusiones

El elevado abandono por parte de los estudiantes que cursan el Grado de Estadística puede ser una preocupación para la Dirección del Grado, pues supone el fracaso académico de una parte importante de los estudiantes que acceden a él. Ello implica también que la universidad pública ha hecho un esfuerzo presupuestario en la subvención de los estudios de estudiantes que no logran sacarlos adelante con éxito. Al mismo tiempo, otros estudiantes quizá más capacitados o motivados no han tenido plaza para acceder a la titulación.

De acuerdo con el análisis del modelo de regresión logística utilizado, se ha observado que hay cuatro factores que influyen en la tasa de abandono del Grado de Estadística. Estos factores explicativos influyen de la siguiente manera:

- El curso inicio hace referencia al año que el estudiante inicia sus estudios. Los estudiantes que empezaron el año 2010 tienen tendencia a abandonar más que los estudiantes que empezaron el año 2009. Concretamente, la probabilidad de abandono es 3,4404 veces mayor. Hay que notar que esta irregularidad puede estar influenciada por tratarse de los años de inicio del Grado de Estadística, en los que el volumen de acceso de estudiantes fue reducido. Si bien no existe significación estadística, el análisis descriptivo muestra que las tasas de abandono siguen una tendencia decreciente, solo ligeramente interrumpida el año 2013.
- En los estudiantes que alcanzan unas notas de acceso superiores al grado, la probabilidad de abandono se reduce en un 62,15% en los estudiantes que obtienen una nota entre (7,9] y un 88,37% en los estudiantes que logran notas superiores a 9, ambos respecto a los estudiantes que tienen una nota entre un 5 y un 7. Esta conclusión es bastante coherente ya que a medida que sube la nota de acceso, el rendimiento académico esperado del estudiante es mayor, y la probabilidad del suceso de abandono es menor.
- La vía de acceso informa que aquellos estudiantes que han llegado a la universidad a través de un ciclo formativo de grado superior, tienen una probabilidad de abandono 4,0403 veces más alta que los estudiantes que han accedido por las pruebas de acceso a la universidad (PAU). A priori, los estudiantes que han realizado las PAU están más preparados en las asignaturas que tienen una relación más estrecha con esta carrera universitaria. Este resultado resulta también coherente, pues el grado de Estadística tiene un grado de complejidad matemática y de abstracción difícil de alcanzar sin estudios previos con un cierto nivel de matemáticas.
- Si los padres no poseen estudios universitarios, la probabilidad de abandono de los estudiantes se reduce en un 52,03% respecto a los estudiantes con padres que poseen dichos estudios. Así, los estudiantes que sus padres no tengan estudios universitarios, quizás alentados por estos, suelen abandonar menos el Grado de Estadística.

Si bien la calidad de ajuste del modelo estimado es aceptable, con un porcentaje de clasificación correcta alrededor del 70%, sería interesante plantear a qué se debe que haya estudiantes que el modelo clasifica como abandonos y que realmente no abandonan. Si bien no disponemos de información para encontrar las causas, estos casos revelan que, a pesar de las condiciones desfavorables de acceso al grado, es posible concluirlo con éxito. Ello debe alentar, por ejemplo, a los estudiantes que provienen de ciclos formativos, o con una nota de acceso limitada a seguir adelante, sabiendo que pueden finalizar sus estudios con éxito.

Igualmente, los responsables académicos deberían prestar atención a los estudiantes que el modelo clasifica como no abandonos, y que realmente abandonan. Dichos estudiantes tienen las condiciones de

acceso óptimas para superar sus estudios y, sin embargo, no lo logran. La ayuda de la tutorización puede ser de vital importancia para este perfil. Dado que a priori no es conocido si el estudiante abandonará o no, los tutores deben estar atentos a los primeros síntomas de bajo rendimiento para ayudar al estudiante a poner remedio a tiempo.

Como limitaciones de este estudio, hay que decir que la reciente creación del Grado de Estadística, que ahora cumple 8 años, tiene como consecuencia que el número de alumnos analizados no es muy elevado. Posiblemente, con una base de datos más amplia y con información de más estudiantes a lo largo del tiempo, la calidad del modelo podría mejorar. Por otra parte, el modelo también mejoraría con la recopilación de otras variables explicativas que no han estado disponibles para este estudio y que pueden influir en el suceso de abandono del Grado de Estadística.

Referencias

Álvarez Pérez, P.R., Cabrera Pérez, L., González Afonso, M.C., Bethencourt Benítez, J.T. (2006) Causas del abandono y prolongación de los estudios universitarios. *Paradigma*, 27, 1, pp. 1-22.

Bernardo, A., Esteban, M., Fernández, E., Cervero, A., Tuero, E., Solano, P. (2016) Comparison of personal, social and academic variables related to university drop-out and persistence. *Frontiers in Psychology*, 7(1610), pp. 1-9.

Bozdogan, H. (1987) Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), pp. 345-370.

Brennan, J. (2008) Higher education and social change. *Higher Education*, 56, pp. 381-393.

Lassibille, G., Navarro, L., Aguilar, I., de la O, C. (2001) Youth Transition from School to Work in Spain. *Economics of Education Review*, 20, pp. 139-149.

Park, L.E.; Crocker, J.; Kiefer, A.K. (2007) Contingencies of self-worth, academic failure, and goal pursuit. *Personality and Social Psychology Bulletin*, 33(11), pp. 1503-1517.

Peng, C.J.; Lee, K.L.; Ingersoll, G.M. (2002) An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), pp. 3-14.

Rubb, S. (2003) Overeducation in the labour market: A comment and re-analysis of a meta-analysis. *Economics of Education Review*, 22, pp. 621-629.