Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. **Rect@**
Volumen 18 (2017), páginas 177 a 201.
DOI: 10.24309/recta.2017.18.2.05

APLICACIÓN DE ÁRBOLES DE CLASIFICACIÓN A LA DETECCIÓN PRECOZ DE ABANDONO EN LOS ESTUDIOS UNIVERSITARIOS DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

JOSÉ MARÍA ORTIZ LOZANO
jmortiz@comillas.edu
Universidad Pontifica Comillas (ICAI/ICADE)
Alberto Aguilera 23 28015 Madrid

ANTONIO RUA VIEITES
rvieites@comillas.edu
Universidad Pontifica Comillas (ICAI/ICADE)/ Departamento de Métodos Cuantitativos
Alberto Aguilera 23 28015 Madrid

PALOMA BILBAO CALABUIG
pbilbao@comillas.edu
Universidad Pontifica Comillas (ICAI/ICADE)/ Departamento de Gestión Empresarial
Alberto Aguilera 23 28015 Madrid

Recibido (24/02/2017)
Revisado (23/10/2017)
Aceptado (23/10/2017)

RESUMEN: El fenómeno de los abandonos en el nivel universitario se produce mayoritariamente en el primer curso, con una media dentro del Sistema Universitario Español del 25%. Las tasas de abandono altas son asociadas, además, con una enseñanza de baja calidad. Con el objeto de ayudar en los procesos de tutorización de estudiantes en la universidad, en el presente trabajo se analiza si es factible determinar en tres momentos: en el momento de la admisión de aquél, del inicio del curso académico y tras la realización de los primeros exámenes, un perfil del estudiante que termina presentando bajo rendimiento académico durante su primer año de estudios. Este trabajo se ha llevado a cabo mediante la aplicación de árboles de clasificación basado en los algoritmos QUEST y CART, sobre una muestra de 844 estudiantes de nuevo ingreso del Grado en Administración y Dirección de Empresas de la Universidad Pontificia Comillas. Se ha obtenido un 56% de tasa de acierto en la clasificación de estudiantes que terminan presentando bajo rendimiento académico, basada en la información disponible al finalizar el primer cuatrimestre.

Palabras clave: Abandono escolar, Bajo rendimiento académico, Árboles de clasificación, Sistema universitario español.

ABSTRACT: Dropouts in university occur mainly in the first academic year, with an average for Spain of 25%. High dropout rates lead to prejudice against educational institutions, it harms their reputation in terms of low quality. In order to help the processes of tutoring students in the university, our work analyzes if it is feasible to get a profile of the student who is at risk of having a low academic performance in his first year in three different moments: when the admission takes place, at the beginning of the academic year, and after the first examinations. This study has used the classification tree technique based on the CART and QUEST algorithms and has used data from 844 first year students enrolled in the Business Administration Licentiate Degree at the Universidad Pontificia Comillas. We have obtained a 56% percentage of correct classified observations for those students who end up presenting low academic performance, with the information available at the end of the first semester.*Keywords*: Student withdrawing, Low academic performance, Classification trees, Spanish university system.

Keywords: Student withdrawing, low academic performance, classification trees, Spanish university system.

1. Introducción

El abandono de los estudiantes universitarios en cualquier tramo universitario en general y, especialmente, durante el transcurso de su primer año de estudios es un problema importante. Entre las situaciones que comúnmente se asocian con el abandono universitario se encuentran: los abandonos obligados por incumplimiento de los requisitos académicos de permanencia en los estudios, dejar los estudios para iniciar otros en la misma u otra institución, abandonar para adquirir formación de índole no universitaria, o incorporarse al mercado laboral, entre otras posibilidades (Cabrera *et al*., 2006).

En España, existe un indicador oficial que da cuenta de qué se considera abandono, la tasa de abandono formulada por la Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA). Esta tasa muestra año a año el porcentaje ligado al abandono en los tres primeros cursos de permanencia de los estudiantes en un mismo plan de estudios. Por tanto, "valores elevados de este indicador debieran motivar un análisis de dónde (y por qué) se produce este abandono, para poder adoptar las medidas correctoras oportunas". En este sentido, los abandonos son los responsables de elevar los indicadores de fracaso y no la tasa de éxito de los alumnos que se presentan a los exámenes (Cernuda del Río *et al*., 2007).

Altas tasas de abandono y/o de bajo rendimiento académico son indicadores de baja calidad de la enseñanza proporcionada por la institución, o directamente de la propia institución, (Berge y Huang, 2004; Cabrera *et al*., 2006; Lykourentzou *et al*., 2009; O'keeffe, 2013; Yorke, 2004), pues se asume que la universidad no ha puesto los medios necesarios y la ayuda suficiente al estudiante con dificultades (Cabrera *et al*., 2006), así como una erosión de la imagen de marca de la institución (Lykourentzou *et al*., 2009) y una infrautilización de los recursos (Tinto, 1975, citado por Yasmin, 2013).

El fenómeno de los abandonos en el nivel universitario se produce mayoritariamente en el primer curso (Bartual y Poblet, 2009; Cabrera *et al*., 2006; Corominas, 2001; Murtaugh *et al*., 1999). Para hacerse una idea de la magnitud del problema, y aunque dependiendo del sistema educativo, tipo de estudios cursado y modalidad del tipo de enseñanza (presencial, semipresencial o a distancia), la literatura considera que hay una media de abandonos en primer curso, dentro del Sistema Universitario Español (SUE), del 25% de los alumnos; llegando en algunas titulaciones hasta el 50% (Cabrera *et al*., 2006).

Existe extensa literatura sobre el fenómeno de los abandonos y/o del bajo rendimiento académico en el nivel universitario, así como de sus causas, factores asociados y variables que han de ser estudiadas para su inferencia. Entre ellos, por citar algunos, se pueden consultar los trabajos de Adam y Gaither (2005), Boyles (2000), Cabrera *et al*., (2006), Cernuda del Río *et al*., (2007), Corominas (2001), Hagedorn (2005), Herzog (2006), Mohammadi (1994), Murtaugh *et al*., (1999), O'keeffe (2013), Rodríguez *et al*. (2004), Texas State Higher Education Coordinating Board (2004), Tinto (2006, 2007), Vivian (2005), Wild y Ebbers (2002) y Yasmin (2013), así como la revisión de trabajos relacionados realizado por Adam y Gaither (2005). En todo caso, cabe referir que en sus inicios la literatura se apoyaba fundamentalmente en el modelo longitudinal de Tinto (1975, 1988) como base para determinar las razones por las que un alumno abandonaba; en dicho modelo se considera la influencia de la integración social y académica del alumno en la institución, las características y capacidades de éste y sus objetivos antes de ingresar y durante el transcurso de sus estudios.

No obstante, si bien las investigaciones académicas relacionadas con el fenómeno de los abandonos y/o del bajo rendimiento académico en el nivel universitario inicialmente se realizaban dentro de un contexto de causa efecto, con una parte descriptiva de las causas y prescriptiva para las soluciones, donde se señalaba que el estudio de los abandonos es un problema complejo y multidimensional, las últimas investigaciones están cambiando el foco y los esfuerzos de las acciones del estudiante a investigar aspectos asociados a la institución e intervenciones de ésta para evitar los abandonos. En este sentido, se han encontrado fundamentalmente tres tipos de factores que intervienen en la decisión para que un alumno abandone, sobre los que se pueden arbitrar medidas de distinta naturaleza para intentar reducirlos: 1) factores sociales o circunstanciales, 2) factores organizativos de la institución, características y procesos de la institución y factores económicos, y 3) factores cognitivos de los alumnos (Berge y Huang, 2004; Swail, 2004).

Los factores institucionales de cualquier Universidad o Institución de Enseñanza Superior (UIES) española representan aquello sobre los que puede influir el *management* de la universidad. Así, la gestión de los abandonos y/o del bajo rendimiento académico que realice la UIES se encuentra vinculado con la aplicación de principios de calidad y promoción de los conceptos fundamentales de excelencia dentro de la organización. En esta línea, Cabrera *et al*., (2006) afirman que la actuación sobre los factores institucionales para reducir el abandono en el nivel universitario, ha sido desarrollada en los últimos años en otros países de la Unión Europea y Norte-América, pero no en España. El mismo autor también destaca alguna de las medidas implementadas: 1) la efectividad de acciones destinadas al reclutamiento, 2) la orientación antes de la elección de estudios sobre las características de las titulaciones, y 3) las acciones de tutoría personalizada y programas de información sobre los alumnos admitidos. Las dos primeras tienen que ver directamente con los procesos de admisión y han de ser desarrolladas antes de la admisión de los alumnos; mientras, la tercera actúa sobre aspectos en los alumnos de primer año. Los programas de tutorización y medidas de asesoramiento de alumnos de primer año demuestran su impacto positivo en la prevención del abandono y/o del bajo rendimiento académico en el nivel universitario (Cabrera *et al*., 2006; O'keeffe, 2013), y para aumentar su eficiencia se pueden arbitrar medidas que permitan inferirlo (Herzog, 2006). Este mismo autor afirma que si bien los beneficios de identificar estudiantes en situación de riesgo, para que sean seguidos en programas de tutorización, son difíciles de concretar, pues el éxito último depende de la eficacia de los programas llevados a cabo para facilitar la finalización de los estudios, en todo caso aumenta las probabilidades de éxito.

En este contexto, se significa que con los alumnos de primer año, existe información de las características de los propios estudiantes, de su rendimiento académico preuniversitario y durante el transcurso del primer semestre de estudios, del colegio de procedencia o de su formación académica anterior, que puede ser considerada –al formar parte de la gestión de los factores institucionales de la UIES- para tratar de aproximar las características de un perfil de alumno con mayor probabilidad de cursar baja o presentar un bajo rendimiento académico con margen suficiente para que la organización ejecute sobre ellos programas de tutorización y medidas de asesoramiento específicos con el propósito de reducir su ocurrencia.

Así, el objetivo principal del presente trabajo es estudiar si del análisis de la información de los solicitantes es factible determinar un perfil del estudiante con riesgo de presentar bajo rendimiento académico (incluyendo el abandono) en el momento de la admisión de aquél, del inicio del curso académico y tras la realización de los primeros exámenes, que sirva de ayuda, por ejemplo, en los procesos de tutorización de estudiantes en la universidad. Dicho estudio se llevará a cabo desde un enfoque multivariante y mediante la aplicación de un modelo basado en un análisis de segmentación con árboles de clasificación.

Para dar respuesta al objetivo propuesto, en el epígrafe 2 se dará cuenta del marco teórico relacionado con las diferentes variables y técnicas utilizadas para inferir el rendimiento académico en el momento de la admisión de los alumnos. El epígrafe 3 se centrará profusamente en el desarrollo teórico de los árboles de clasificación. En el epígrafe 4 se aplicará el método seleccionado a una muestra para pasar, en los últimos dos epígrafes, 5 y 6, a la descripción de los resultados y conclusiones respectivamente.

2. Marco teórico

En el presente epígrafe se presentan, las variables y técnicas más utilizadas con objeto de clasificar a aquellos alumnos que presentan una mayor probabilidad de cursar baja o tener rendimiento académico bajo durante el primer año de sus estudios, prestando especial atención al análisis de segmentación mediante la técnica de los árboles de clasificación.

2.1. *Variables utilizadas para inferir el rendimiento académico en el momento de la admisión de los alumnos*

Se presenta a continuación una breve síntesis de la literatura que ha estudiado la predicción del rendimiento académico referidas al primer año o curso académico (Alcover *et al*., 2007; García y San Segundo, 2001; Guisande *et al*., 2006; Rodríguez y Coello, 2008; Rodríguez *et al*., 2004; Shulruf *et al*., 2008), o al finalizar un ciclo de estudios completo (Carrión, 2002; Goberna y López, 1987; Peña y Sánchez, 2005). Se obtienen tasas predictivas menores cuando en el momento de la admisión se infiere el rendimiento académico del ciclo de estudios completo que cuando se estudia la capacidad predictiva sobre el rendimiento académico para el primer curso académico, primer año, o asignaturas concretas del primer año (Rúa y Kennedy, 2003), aunque ambas medidas están altamente correlacionadas de manera positiva (Díaz y Toloza, 2007; Murtaugh *et al*., 1999).

En cuanto a la medida del rendimiento académico que dan las autoridades con competencias en materia educativa en España, se destacan dos tipos: a) el cociente entre créditos aprobados y créditos matriculados, denominado tasa de rendimiento, utilizado por la ANECA para la acreditación de los títulos universitarios de Grado implantados; b) las notas medias utilizadas para la mayoría de los sistemas de becas, donde se utiliza una nota media ponderada por créditos al finalizar el año académico.

Entre las principales variables que se han utilizado para inferir el posterior rendimiento académico que los alumnos admitidos en estudios universitarios presentan posteriormente, destacan, por su frecuencia, el rendimiento académico preuniversitario y, en los países que exista, como ha sido el caso de España, la Prueba de Acceso a la Universidad (PAU) (Alcover *et al*., 2007; Betts y Morell, 1999; Cortés Flores, 2008; García Jiménez *et al*., 2000; García y San Segundo, 2001; Gimeno *et al*., 2003; Marcenaro y Navarro, 2007; Rodríguez y Coello, 2008; Shulruf *et al*., 2008; Vélez y Roa, 2005), no presentándose consenso para determinar cuál de ellas presenta mayor poder predictivo. Si bien, uno de los principales problemas que presentaba la PAU es consecuencia de la corrección de las mismas por tribunales distintos (Cuxart *et al*., 1997; Escudero (1987) y Goberna y López (1987)).

Otro factor que puede ser considerado conforme a la normativa vigente que establece la normativa básica de los procedimientos de admisión a las enseñanzas universitarias oficiales de Grado es la modalidad de Bachillerato realizada. La literatura consultada, aun con distintas modalidades, ya ha estudiado la influencia de esta variable para determinar posteriormente el rendimiento académico de los alumnos. Así, los estudiantes de Facultades de Ciencias Económicas y Empresariales, que han estudiado matemáticas de la modalidad de Ciencias obtienen mejor rendimiento académico que quienes realizaron matemáticas en la modalidad de Ciencias Sociales (Castellanos Val *et al*., 1998). Otros autores que han encontrado relaciones entre la modalidad cursada durante el Bachillerato y el rendimiento académico en la Universidad son Escudero (1984), Herrera *et al*. (1999) y Rúa y Kennedy (2003).

2.2. *Técnicas estadísticas clásicas utilizadas para inferir el rendimiento académico en el momento de la admisión de los alumnos*

Existen fundamentalmente dos tipos de técnicas estadísticas utilizadas para inferir el rendimiento académico universitario que posteriormente logran los alumnos que son admitidos (Mafokozi *et al*., 2001):

- Modelos experimentales, que estudian si la variable modificada durante la realización del experimento influye en el rendimiento académico. En este caso, las técnicas de análisis de datos más utilizadas se basarían en contrastes de medias entre los grupos con la variable influida y los grupos que sirven de control.
- Modelos correlacionales, donde se puede diferenciar entre los modelos correlacionales puros y los modelos correlacionales con propósito de causalidad. Dentro de este grupo, si bien existen autores que han preferido realizar análisis discriminantes (García Llamas, 1986) o análisis de conglomerados (Rúa Vieytes *et al*., 2003), que suelen completarse con otro tipo de análisis

multivariante posterior, destacan por su frecuencia de aplicación las técnicas de regresión lineal múltiple (Cortés, 2008; Cuxart *et al*., 1997; Gallestey *et al*., 2004; García y San Segundo, 2001; Gimeno *et al*., 2003; Rúa y González, 2004; Vélez y Roa, 2005) y logística (Bartual y Poblet, 2009; Carrión, 2002; Díaz y Toloza, 2007; Peña y Sánchez, 2005; Press y Wilson, 1978). Dentro de este grupo también se encuentran los árboles de clasificación, los cuales suponen una alternativa a las técnicas estadísticas más clásicas, como la regresión múltiple, los análisis ANOVA, regresión logística, análisis discriminante y modelos de supervivencia (De'ath y Fabricius, 2000). Los árboles de clasificación, además, parecen obtener mejores tasas predictivas que el resto de técnicas de minería de datos cuando se utiliza información mayoritariamente de tipo categórico (Adam y Gaither, 2005; Herzog, 2006; Jing, 2002; Yasmin, 2013), como es el caso de las variables que utiliza la presente investigación.

Con el objeto de investigar qué variables y categorías de éstas clasifican a aquellos alumnos que presentan una mayor probabilidad de cursar baja o tener rendimiento académico bajo durante el primer año de sus estudios, este trabajo utiliza fundamentalmente un análisis de segmentación mediante la técnica de los árboles de clasificación.

2.3. *Aplicación de los árboles de clasificación para predecir el abandono escolar*

A grandes rasgos, los árboles de clasificación son una técnica estadística que permite explicar la variación de una variable con respuesta única. Para ello, se utiliza la realización de una continua división de datos, a través de un proceso secuencial descendente, en grupos homogéneos, exhaustivos y mutuamente excluyentes. Sus principales aplicaciones son la exploración, descripción y predicción de patrones y procesos que emplean y combinan las variables explicativas (Mercado, 2007; Yasmin, 2013).

Los árboles de clasificación utilizan algoritmos de segmentación y se enmarcan dentro de las técnicas referidas como minería de datos (Baradwaj y Pal, 2012; Berlanga *et al*., 2013; Jing, 2002; Mercado, 2012; Pérez y Santín, 2007; Rojo, 2006; Superby *et al*., 2006; Vandamme *et al*., 2007).

La principal justificación para la aplicación de esta técnica es que la misma ha demostrado su utilidad descriptiva, exploratoria y explicativa, así como su aplicación para la clasificación de datos, en el ámbito del rendimiento académico (Angúlo y Sergio, 2012; Baradwaj y Pal, 2012; Campbell *et al*., 2007; Dekker *et al*., 2009; Gallestey *et al*., 2004; Herzog, 2006; Jing, 2002; Kumar y Vijayalakshmi, 2011; Mercado, 2007, 2012; Nghe *et al*., 2007; Porcel *et al*., 2009; Quadri y Kalyankar, 2010; Ramaswami y Bhaskaran, 2010; Rojo, 2006; Superby *et al*., 2006; Vandamme *et al*., 2007).

Son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclan datos categóricos y numéricos (Alcover *et al*., 2007), y éste es el escenario que se presenta en el propósito de aproximar las características que presentan aquellos alumnos de nuevo ingreso que tienen mayor probabilidad a priori de cursar baja, o presentar un rendimiento académico bajo, en su primer año de estudios. Además, una ventaja especialmente valorada para su aplicación por responsables de los procesos de admisión en una universidad, como pueden ser los distintos tipos de gestores o personal de administración y servicios, es que no es preciso una habilidad analítica excepcional para afinar un árbol de decisión (Rojo, 2006).

Los árboles de clasificación difieren notablemente respecto de otras técnicas de clasificación multivariable, como pueden ser el análisis discriminante, el análisis de regresión Logit, el análisis factorial o el análisis de conglomerados. Las tres grandes diferencias que hacen singulares en su conjunto a los árboles de clasificación son (Mercado, 2007): 1) trabajan con la variable dependiente en su estado original, sin transformación alguna, aspecto que facilita la interpretación de los datos; 2) el procedimiento de segmentación utilizado es descendente, dividiendo la muestra sucesivamente y detectando las interacciones automáticamente; y 3) permite la obtención de una tabla de clasificación con el porcentaje de acierto y su representación gráfica con forma arbórea. Las principales ventajas en el uso de los árboles de clasificación respecto de otras técnicas más clásicas son (Rojo, 2006; Yasmin, 2013): Proporcionan

flexibilidad en manejar un amplio rango de tipos de repuesta, entre ellas, numéricas continuas, categóricas, listas y/o datos de supervivencia; ofrecen invariabilidad a transformaciones de las variables explicativas; aportan facilidad y robustez del sistema; son fáciles de interpretar; tienen capacidad de manejar datos vacíos, tanto en variables explicativas como en la variable dependiente.

Dado lo relativamente novedoso de la aplicación de este tipo de técnica respecto de otras más clásicas, ya mencionadas, en la inferencia del rendimiento académico para aproximar características de los alumnos con riesgo alto de cursar baja, o presentar bajo rendimiento académico, en su primer año de estudios universitarios, en el ámbito del SUE, y especialmente de alguno de los algoritmos utilizados (en concreto, el algoritmo QUEST[*]), en el siguiente epígrafe se expondrán, sin entrar en demasiados ambages estadísticos, los principales fundamentos utilizados por los árboles de clasificación en referencia al alcance de aplicación de los mismos en este trabajo de investigación.

3. Árboles de clasificación

En primer lugar, se introduce la definición de un conjunto de conceptos utilizados con frecuencia en los árboles de clasificación. Así, se define un árbol con k-hijos como un grafo formado por nodos y aristas, que ha de cumplir lo siguiente: existe un único nodo que no tiene padre, denominado nodo raíz; cualquier nodo distinto al nodo raíz tiene un único nodo padre; cualquier nodo tiene ninguno o varios hijos. Los nodos que no tienen hijos se llaman nodos terminales u hojas.

A modo de ejemplo, en la figura 1 se presenta un grafo que muestra la estructura de un árbol de clasificación. El mismo tiene dos niveles de profundidad y 3 hojas o nodos terminales que han sido segmentados en su primer nivel por una variable cuantitativa X, en función de un determinado valor z, y en su segundo nivel por una variable cualitativa Y que toma los valores a o b.

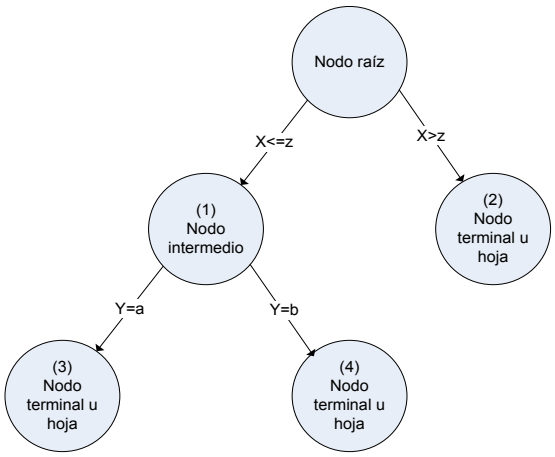

Figura 1. Ejemplo de grafo que representa un árbol de clasificación

El nodo raíz representa la variable a segmentar. Cada arista determina la división realizada sobre el nodo padre en base al valor o valores de una variable independiente. Los nodos terminales u hojas representan un subconjunto de valores homogéneos sobre la variable de estudio donde el mejor criterio para el pronóstico de aquélla viene determinado por la categoría modal en él.

Dentro del análisis de segmentación existen variantes, en función del algoritmo estadístico empleado para realizar la división y selección de variables independientes. Los más conocidos en la actualidad son los denominados como AID[†], CHAID[‡] y su variante CHAID exhaustivo, CART[§] y QUEST. En la Tabla 1 se adelanta un resumen de las principales diferencias entre los algoritmos de segmentación.

[*] QUEST, de las siglas en inglés, *Quick, Unbiased, Efficient Statistical Tree.*

[†] AID, de las siglas en inglés *Automatic Interaction Detection.*

Tabla 1. Comparación de los distintos algoritmos de segmentación. Fuente: Mercado (2007)

Característica	AID	CHAID	CART	QUEST
V. Dependiente	Cuantitativa	Cualitativa	Ambas	Cualitativa
Predictores	Ambos	Cualitativos	Ambos	Ambos
Fusión	Binaria	Múltiple	Binaria	Binaria
Fisión	MCE o F	Significación	Índices de mejora	Significación
Orden	Fusión/Fisión	Fusión/Fisión	Fusión/Fisión	Fisión/Fusión
Tratamiento de casos perdidos	Autónomo	Autónomo	Sustitutos	Sustitutos
Filtros	Asociación	Significación	Mejora	Significación

Dado que el uso de esta técnica estadística en este artículo se plantea para determinar el perfil que presenta el alumno que cursa baja, u obtiene un rendimiento académico bajo, en el primer año de estudios, la variable dependiente es de naturaleza cualitativa. Por ello, no se podrá utilizar AID. Asimismo, dada la presencia de variables independientes de tipo cualitativo y cuantitativo, los algoritmos más adecuados para realizar la segmentación serían CART y QUEST. Si bien es cierto que CHAID podría ser considerado para la obtención de resultados, exigiría una transformación de las variables independientes cuantitativas en ordinales con un bajo número de valores, con el inconveniente de la posible pérdida de información que lleve implícita la transformación (Goicoechea, 2002).

A continuación, se exponen brevemente los principales fundamentos utilizados por los algoritmos CART y QUEST sin entrar en demasiada profundidad de técnica estadística. No obstante, si se quiere profundizar en ellos se recomienda la lectura de Breiman *et al*. (1984) para CART y de Loh y Shih (1997) para QUEST.

La diferencia entre árboles de clasificación y regresión viene determinada por la naturaleza de la variable dependiente. Si ésta es cualitativa se denominan árboles de clasificación (CT[**]), mientras que si es cuantitativa nos estamos refiriendo a árboles de regresión (RT[††]). Lógicamente, el criterio para la división de los grupos es distinto si la variable dependiente es cualitativa, donde se utilizan índices de mejora basados en medidas de impureza, que si la variable dependiente es cuantitativa, donde se utilizarán índices de mejora basados en mínimos cuadrados. En este último caso, existe una gran similitud entre RT y AID.

El origen de estas técnicas de segmentación se atribuye a Breiman, Friedman, Stone, y Olshen (Mercado, 2007), quienes desarrollaron nuevos algoritmos de segmentación, siempre binarias, que son utilizados bajo el acrónimo CART.

Estos algoritmos utilizan medidas de selección basadas en una aproximación de la impureza dentro de los grupos.

[‡] CHAID, de las siglas en inglés *Chi Automatic Interaction Detection.*

[§] CART, de las siglas en inglés *Classification an Regression Trees*.

[**] CT de las siglas en inglés *Classification Trees.*

[††] RT de las siglas en inglés *Regression Trees.*

3.1. *Medidas de selección de CART*

Dada una variable cualitativa Y, ésta puede tomar J valores representado cada uno de ellos por la letra j. Si Y es dividida en T grupos, cada uno de ellos presentará una frecuencia relativa respecto de p(j); como es lógico,

$$\sum_{t=1}^{T} p_t(j) = 1. \quad (1)$$

Una segmentación es completamente homogénea si presenta p(j)=1 en un grupo, mientras que el resto de grupos tiene p(j)=0. Por el contrario, la mayor heterogeneidad se presenta cuando p(j)= 1/J en todos los grupos.

Aproximados los anteriores conceptos estadísticos, se pueden incorporar la definición de dos índices que dan cuenta del poder homogeneizador como criterio para la segmentación de grupos en CR, el índice de Gini y el índice binario.

3.1.1. Índice de Gini

Se introduce el índice de diversidad o de impureza de un determinado grupo como la suma de la probabilidad de la coocurrencia de dos valores distintos. Así, por ejemplo, para un grupo con dos valores, el valor de la diversidad será dos veces el producto de p_1 por p_2, pues es $p_1xp_2 + p_2xp_1$; para un grupo con tres valores, el valor de la diversidad será dos veces las sumas del producto de p_1 por p_2, p_1 por p_3 y p_2 por p_3. En consecuencia, el mayor valor de diversidad se alcanzará dentro de un grupo cuando las p(j) sean 1/J, alcanzando el valor de (J-1)/J.

El índice de Gini se generaliza dentro de un grupo t mediante la siguiente fórmula, dada por dos expresiones equivalentes:

$$i(t) = \sum_{i\neq 1}^{J(J-1)} p(j)p(i) = 1 - \sum_{j=1}^{J} p(j)^2 \quad (2)$$

Una propiedad de este índice es que cuando un grupo se segmenta la suma ponderada del índice de diversidad antes de la segmentación siempre será mayor o igual a la suma ponderada de los índices de diversidad de los segmentos generados. En el caso de la segmentación en dos grupos –o binaria-, que es la que es empleada por el algoritmo CR, se puede expresar matemáticamente mediante la siguiente formula.

$$i(t) \geq p_L i(t_L) + p_R i(t_R) \quad (3)$$

Por tanto, puede determinarse el poder de homogeneización de la segmentación realizada de cualquier clasificador (s) para un grupo (t) mediante la diferencia del índice de diversidad antes de la división (i(t)) y la suma ponderada de los índices de diversidad de los grupos segmentados a la izquierda (L) y a la derecha (R) respectivamente.

$$\Phi(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (4)$$

3.1.2. Índice binario

En el caso en el que la variable dependiente sea dicotómica, como es el caso de esta investigación, el índice binario se define como una media al cuadrado de las diferencias absolutas de los porcentajes de cada grupo segmentado por el producto del peso de los grupos segmentados.

$$\Phi(s,t) = \left[\frac{\sum_{j=1}^{2}|p(j|t_L) - p(j|t_R)|}{2}\right]^2 p_L p_R \quad (5)$$

En consecuencia, la utilización del índice binario favorece la división de grupos donde se produzcan diferencias dentro de los grupos altas y tamaños de los grupos similares, pues $p_L p_R$ alcanza su máximo cuando la proporción de casos en ambos grupos es 0,5.

Se introducen a continuación el resto de criterios básicos en los que se sustenta CART: la fusión de categorías de los clasificadores, la selección de la mejor variable, el procedimiento de realización de nuevas particiones, así como las reglas de parada en la construcción del árbol.

3.1.3. Fusión de categorías, dicotomización y selección de variable

Dado que CART siempre propone una división binaria, el criterio es sencillo. En el caso de variables independientes cualitativas se trata de hallar aquella división binaria de categorías, dentro de cada variable, que obtenga el mayor índice de Gini o, en su caso, índice binario.

Como se puede anticipar, en el caso de que la variable independiente cualitativa que se esté considerando sea nominal, el número de divisiones binarias posibles puede ser muy alto, con el consiguiente aumento del tiempo de procesamiento de cálculo. Téngase en cuenta que una variable cualitativa nominal con c categorías tendría un número de particiones binarias posibles igual a 2^{c-1}-1.

En el caso de que la variable independiente sea cuantitativa no hay mayor problema en ir realizado sucesivas segmentaciones hasta determinar el valor de corte que obtenga el mayor índice de Gini/binario.

Para determinar la variable que habrá de producir la segmentación se han de comparar los índices de Gini/binarios obtenidos en la división binaria de todas las variables referida anteriormente y aquella que obtenga el mayor índice deberá producir la siguiente segmentación.

3.1.4. Nuevas particiones y reglas de parada

Con la segmentación producida en el grupo anterior se debe repetir el proceso nuevamente, incluyendo todas las variables independientes. Esto permite que una misma variable pueda producir distintas segmentaciones de forma sucesiva, siempre que cumpla con el criterio de obtener el mejor índice de mejora. Este procedimiento se deberá iterar sucesivamente, obteniendo nuevas particiones, hasta que se alcance una de las reglas de parada establecidas.

Existen tres criterios fundamentales para establecer reglas de parada, la basada en un índice de mejora mínimo, número mínimo de casos de los grupos y profundidad máxima del árbol.

El primer criterio de parada consiste en no tener segmentaciones posibles que no superen un índice de mejora mínimo. Es norma establecer un índice de mejora bajo (0,0001) para luego proceder a realizar una poda del árbol, procedimiento que se explica más adelante.

El segundo criterio de parada se refiere a la existencia de límites de mínimo de casos en los grupos para poder seguir segmentando. Existen dos tipos de límite: límite para el nodo filial (nodo tras la segmentación) y límite para el nodo parental (nodo antes de la segmentación). En el caso del límite que aplicaría a un nodo filial, dicho límite impedirá que se produzca la segmentación si ésta conlleva la creación de un grupo con un número de casos inferior a dicho límite. En el caso del límite que aplica sobre un nodo parental, impedirá que éste pueda segmentarse si el número de casos de dicho nodo parental es inferior al límite establecido. Como regla general, se establece como referencia, 50 casos para un nodo filial y 100 para un nodo parental (Mercado, 2007).

El tercer criterio de parada consiste en limitar la profundidad del árbol, para evitar que éste se vuelva demasiado frondoso y conlleve un aumento de la dificultad para la interpretación del árbol.

Otros criterios de parada menos frecuentes pueden ser obtener valores idénticos de índices de mejora u obtener nodos puros (todos sus casos tienen el mismo valor).

3.2. ***Algoritmo QUEST***

La principal novedad que aporta el algoritmo QUEST respecto de CART, así como también respecto de AID y CHAID, es que invierte el proceso de construcción del árbol. En lugar de empezar con la fusión de categorías dentro de cada variable y después proceder a la sección de variable, QUEST determina en primer lugar la mejor variable y después procede a la determinación del mejor corte dentro de aquélla. Este procedimiento trata de subsanar una de las mayores debilidades de AID, CHAID y CART, el problema de que se favorezca la selección de variables con mayor número de categorías respecto al resto, aspecto que va en detrimento de la interpretación del árbol (Loh y Shih, 1997).

3.2.1. Selección de variable segmentadora

Un aspecto importante a considerar es que QUEST trabaja necesariamente con variables dependientes cualitativas. En este contexto, QUEST propone para la selección de la variable la ejecución de pruebas de significación entre la variable dependiente y las variables independientes. Así, si la variable independiente es cualitativa nominal utiliza el estadístico $\chi 2$ de Pearson, con número de grados de libertad igual al número de categorías menos una de la variable independiente. Mientras que, si la variable independiente es de intervalo o cuantitativa utiliza el estadístico F con grados de libertad en el numerador igual al número de categorías de la variable dependiente (J) menos uno y grados de libertado en el denominador n-J.

En el supuesto de que ninguna variable se presente significativa, en el caso de predictores cuantitativos se aplica la prueba de Levene basada en las diferencias absolutas de los valores de cada variable respecto de la media de su nodo, siendo ésta una prueba para determinar si las varianzas de la variable predictora son homogéneas en cada una de las categorías de la variable dependiente. Si resultase que la referida prueba arroja que las varianzas no son homogéneas, y siempre que no exista ninguna variable significativa, se utilizaría esta variable para realizar la segmentación.

3.2.2. Determinación del punto de corte para la división binaria

Una vez determinada en el paso anterior la variable que habrá de realizar la segmentación, el procedimiento para la determinación del punto de corte que el procedimiento QUEST utiliza sigue los siguientes pasos: 1) Si la variable predictor elegida es cualitativa nominal ha de convertirse en una variable continua, con valores entre -1 y +1, utilizando para ello un análisis discriminante que transformará cada valor de la variable cualitativa a su puntuación discriminante. Este procedimiento recibe el nombre de CRIMCOORD y fue desarrollado en origen por Gnanadesikan (1977); 2) En el caso de que la variable dependiente tenga más de dos valores y puesto que en este punto la variable predictor solo puede ser cualitativa ordinal o continua, ha de utilizarse un análisis de conglomerados de dos medias (Hartigan y Wong, 1979) para que la variable dependiente se convierta en dicotómica; 3) Dado que la variable dependiente es dicotómica, bien porque lo sea en origen, bien porque se haya transformado a ella en el paso anterior, QUEST lleva a cabo un análisis discriminante cuadrático que determinará el valor de la variable predictor que dividirá a la misma en dos segmentos.

3.2.3. Evaluación de la bondad de la clasificación

Partiendo de que cada nodo terminal se tiene que utilizar para clasificar cada uno de los casos ahí presentes respecto de la variable dependiente, el mejor criterio para realizar un pronóstico en un nodo terminal es utilizar la categoría modal del mismo (Mercado, 2007). No obstante, existirá en todo caso un número de casos mal clasificados.

Así, por tanto, se puede aproximar una medida del riesgo del árbol R(T) como el cociente entre la suma del número de casos m_t mal clasificados en todos los nodos terminales ($\overline{T}$) del árbol y el número total de casos (N).

$$R(T) = \frac{\sum_{t=1}^{|\overline{T}|} m_t}{N} \tag{6}$$

Lógicamente cuanto menor sea el riesgo del árbol mejor será la bondad de la clasificación.

Esta medida del riesgo también puede extenderse si existen costes distintos en función de la categoría mal clasificada en la variable dependiente. Es decir, puede suceder que exista un coste determinado por clasificar incorrectamente un caso como i cuando realmente es j. Expresado matemáticamente como:

$$C(i|j) = \begin{cases} 0, & si\ i = j \\ \geq 0, & si\ i \neq j \end{cases} \tag{7}$$

En tal caso, la estimación del riesgo del árbol quedaría:

$$R(T) = \sum_{t=1}^{|\bar{T}|}\left[min \sum_{j=1}^{J} p(j|t)C(i|j)\right]p(t) \quad (8)$$

donde $p(j|t)$ representa el porcentaje de la variable dependiente presente en el nodo t y p(t) la probabilidad de ocurrencia del nodo terminal t.

3.3.4. Método de validación cruzada para el cálculo de la estimación del riesgo

Para propósitos de causalidad en la utilización de los árboles de clasificación Breiman *et al*. (1984) propuso el método de validación cruzada para el cálculo de la estimación de riesgo, especialmente indicado en muestras pequeñas.

El procedimiento consiste en dividir de forma aleatoria la muestra (compuesta por los N casos) en C conglomerados independientes entre sí, cada uno con un tamaño de N/C casos. A continuación se repite el siguiente procedimiento N veces de forma sucesiva: se construye un árbol (que recibe el nombre de árbol de entrenamiento) tomando los valores de todos los conglomerados menos uno, esto es N((c-1)/c) casos y el resultado se aplica sobre el conglomerado que ha quedado fuera en el paso anterior (a este árbol se le denomina árbol de comprobación).

Así, por ejemplo, si se divide la muestra en 5 conglomerados, el primer árbol utilizará los casos de los conglomerados 1 a 4 y el resultado lo aplicará sobre el conglomerado 5 obteniendo en consecuencia un número de casos mal clasificados, con su correspondiente estimación de riesgo; el segundo árbol construirá la muestra de entrenamiento con los casos de los conglomerados 1,3,4 y 5, y lo aplicará sobre el conglomerado 2 obteniendo un número de casos mal clasificados con su correspondiente estimación de riesgo. El procedimiento se repetirá hasta obtener las cinco estimaciones de riesgo. Para obtener el estimador final del riesgo bastará obtener una media de las estimaciones de riesgo individuales.

En cualquier caso, conviene destacar que Breiman *et al*., (1984) señalan 10 como número óptimo y adecuado de conglomerados para el sistema por validación cruzada.

3.3.5. Procedimiento para la Poda de los árboles

Con anterioridad han sido expuestos procedimientos que limitan el crecimiento de los árboles en base al establecimiento de reglas de parada. No obstante, para evitar el sobreajuste del árbol a los datos, Breiman *et al*., (1984) recomiendan proceder en dos pasos: uno primero de construcción del árbol con el mayor número de nodos puros posibles, para, a continuación, proceder en un segundo paso a realizar la denominada "poda" que elimine aquellas ramas que no aporten una suficiente reducción de la estimación del riesgo.

En este punto, se introducen las siguientes características del riesgo dentro de un árbol:

1) La complejidad de un árbol puede clasificarse en función del número de nodos que aquél contenga. En consecuencia, puede aproximarse una medida del coste complejidad del árbol como:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \quad (9)$$

Notar que si α=0 el coste-complejidad del árbol será igual a la estimación del riesgo del árbol.

2) Los riesgos de los nodos son independientes del nivel en el que se sitúen dentro del árbol.

3) El riesgo de un nodo t, denominado como $R(\{t\})$, siempre será mayor o igual que el riesgo de la rama que salga de él $R(T_t)$. Es decir, $R(\{t\}) \geq R(T_t)$.

De lo anterior, se puede determinar el coste complejidad de un nodo como:

$$R_\alpha(\{t\}) = R(\{t\}) + \alpha \quad (10)$$

Así como el de la rama que sale de cualquier nodo:

$$R_\alpha(T_t) = R(T_t) + \alpha\left|\widetilde{T}_t\right| \tag{11}$$

En consecuencia, existe un valor α, denominado crítico g(t), que determinaría el punto por encima del cual el coste-complejidad de la rama sería mayor que el del nodo y en consecuencia no ha de preservarse la rama. El referido valor crítico g(t) se formula a partir de las ecuaciones anteriores como:

$$g(t) = \frac{R(T_t) - R(\{t\})}{|\widetilde{T}_t| - 1} \tag{12}$$

El procedimiento que se utiliza en la poda del árbol utiliza este valor crítico g(t) eliminando aquellas ramas con g(t)=0 y eliminadas éstas se eliminarán sucesivamente aquellas con el valor mínimo de g(t).

Para determinar donde parar con la poda del árbol, se ha de ir podando el árbol conforme al criterio anteriormente especificado hasta obtener aquél subárbol más pequeño que esté dentro de la máxima diferencia especificada para el riesgo expresado en términos de error típico. El valor del riesgo expresado en error típico se formula como:

$$ET\big(R(T)\big) = \sqrt{\frac{R(T)(1-R(T))}{n}} \tag{13}$$

4. Método

Para investigar el perfil que presentan aquellos alumnos con una mayor probabilidad de cursar baja o tener rendimiento académico bajo durante el primer año de sus estudios, se han llevado a cabo las siguientes acciones:

1. Recolección y tabulación de la información con la que cuenta la universidad del colectivo de alumnos donde se lleva a cabo la investigación.
2. Realización de un estudio exploratorio sobre las variables obtenidas en el paso anterior, con el objeto de realizar una primera aproximación sobre cuáles son las variables que mayor capacidad de discriminación pueden presentar para determinar si un alumno cursará baja al finalizar su primer año de estudios, o presentará un rendimiento académico bajo. El estudio exploratorio ha consistido en la obtención de estadísticos descriptivos y tablas de contingencia, con sus correspondientes contrastes de asociación.
3. Con las variables extraídas del paso anterior, previa comprobación en caso de variables nominales de la existencia de nivel de asociación significativo entre las variables independientes y la variable dependiente, se realiza un análisis de segmentación basado en los árboles de clasificación, que permite aproximar tanto el perfil del alumno que cursa baja al finalizar su primer año de estudios como el perfil del alumno que tiene bajo rendimiento académico en sus estudios durante el primer año.

En este punto, se significa que, aunque la metodología utilizada para realizar la investigación puede ser la misma con independencia del plan de estudios y la universidad donde los estudiantes realicen sus estudios, dado que las características de los alumnos y planes de estudio pueden diferir significativamente entre sí, así como las variables con las que se lleve a cabo el estudio, se ha de escoger necesariamente un colectivo de alumnos perfectamente definido donde realizar la investigación. En otras palabras, es necesario acotar la población de estudio.

4.1. *Participantes*

Para llevar a cabo este análisis se han tomado los datos de 844 alumnos de nuevo ingreso del Grado en Administración y Dirección de Empresas, procedentes de los cursos académicos 2009/2010 al 2014/2015, ambos inclusive, organizado por la Facultad de Ciencias Económicas y Empresariales de la Universidad Pontificia Comillas. El porcentaje de estos alumnos que cursan baja o repiten primer curso en su primer año de estudios es del 12,9%.

A los efectos de este estudio se considera baja el alumno que entra dentro de alguno de los siguientes supuestos: a) Incumplimiento de los requisitos académicos de permanencia, así, el alumno que en su primer año de estudios no supera, entre las convocatorias ordinarias y extraordinaria, al menos el 50% de los créditos no podrá continuar sus estudios; b) El estudiante solicita la baja voluntaria durante su primer año de estudios, habiéndose producido ésta tras haber concurrido al menos a la primera convocatoria del curso académico. Adicionalmente, el primer curso de los estudios del Grado en Administración y Dirección de Empresas tiene carácter selectivo. Por ello, se considera alumno repetidor de primer curso a aquél que, sin haber aprobado todos los créditos, al menos ha superado el equivalente al 50% de los créditos de Grado entre las convocatorias ordinaria y extraordinaria. En ese supuesto los alumnos pueden repetir cursando las asignaturas pendientes, sin poder matricularse en asignaturas de cursos superiores. A este colectivo de alumnos el presente estudio lo considera alumno con rendimiento académico bajo o inferior al esperado.

La suma del colectivo de alumnos repetidores y de alumnos que cursan baja en su primer año de estudios representan el conjunto de estudiantes cuyas características pretenden ser aproximadas con margen de tiempo para poder llevar a cabo medidas específicas de tutorización, seguimiento y apoyo sobre aquéllos que permitan mejorar su rendimiento académico.

4.2. *Instrumentos*

La obtención de la información ha requerido la ejecución de distintas consultas SQL[‡‡] a las bases de datos de Admisiones y de Gestión Académica de la Universidad Pontificia Comillas. Una vez tratada, consolidada y depurada la información se importó al software estadístico SPSS, donde se ejecutaron estadísticos descriptivos, tablas de contingencia con pruebas de nivel de asociación (según corresponda, estadísticos Gamma, Phi y V de Cramer), y distintos análisis de segmentación basados en los árboles de clasificación. Las pruebas estadísticas realizadas siempre han utilizado un nivel de significación del 5%.

En lo referente a las consideraciones tomadas en la construcción de los árboles de clasificación, dada la naturaleza de la variable a segmentar (dicotómica y nominal) se han utilizado los algoritmos CART y QUEST. La elección en la segmentación realizada por cada uno de ellos se ha basado en aquél que presente mayor valor de acierto para la categoría objetivo (cursar baja o presentar un rendimiento académico bajo). Para evitar un sobreajustamiento a los datos se ha utilizado un sistema de poda del árbol basado en tomar el valor de 1 como máxima diferencia del riesgo, medido en errores estándar. Notar que dada la baja frecuencia de valores presentada en la categoría objetivo y que el número de individuos es menor a 900, valor señalado como referencia por Breiman *et al*., (1984) para la adecuada aplicación de los árboles de clasificación, se ha considerado adecuado reducir los valores y considerar un mínimo de 80 casos para poder dividir un nodo y un mínimo de 35 para poder conformar un nodo. Por último, y puesto que el interés radica en maximizar el acierto en la clasificación de aquel alumno que cursa baja o presenta un rendimiento académico bajo en su primer año de estudios, se ha incluido la consideración de un doble coste en clasificar incorrectamente un alumno que es predicho por el modelo como estudiante que continúa cuando en realidad termina cursando baja u obtiene un rendimiento académico bajo.

4.3. *Variables*

Las variables consideradas han sido un total de 26. Éstas se pueden dividir según el momento en el que la Universidad Pontificia Comillas obtiene cada una de ellas en tres grupos: a) información procedente de las pruebas de admisión realizadas por los alumnos de los referidos Grados para el ingreso, b) información acreditativa de los requisitos de acceso o procedente de la estadística universitaria que los alumnos admitidos han de cumplimentar obligatoriamente al formalizar su primera matrícula al inicio del

[‡‡] El lenguaje de consulta estructurado o SQL (por sus siglas en inglés *Structured Query Language*) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en estas

curso académico, y c) información que da cuenta de los primeros resultados obtenidos en el transcurso de los estudios del alumno.

A continuación se exponen las variables consideradas en el estudio llevado a cabo para investigar qué variables y categorías de éstas clasifican a aquellos alumnos que presentan una mayor probabilidad de cursar baja o tener rendimiento académico bajo durante el primer año de sus estudios.

a) Información procedente de las pruebas de admisión realizadas por los alumnos:

1. Sexo del alumno, con posibles valores: 0 mujer; 1, hombre.
2. Media de expediente académico del curso académico realizado tres años antes de la realización de las pruebas de admisión, normalmente 3º de Enseñanza Secundaria Obligatoria (ESO).
3. Media de expediente académico del curso académico realizado dos años antes de la realización de las pruebas de admisión, normalmente 4º ESO.
4. Media de expediente académico del curso académico realizado un año antes de la realización de las pruebas de admisión, normalmente 1º Bachillerato.
5. Media aritmética del expediente académico de los tres cursos anteriores al de la realización de las pruebas de admisión (puntos 2, 3 y 4).
6. Resultado de la nota obtenida en la prueba de admisión de Matemáticas 1.
7. Resultado de la nota obtenida en la prueba de admisión de Matemáticas 2.
8. Resultado de la nota obtenida en la prueba de admisión de Matemáticas 3.
9. Resultado final de la prueba de admisión de matemáticas, en base a las pruebas Matemáticas1, Matemáticas2 y Matemáticas3 (puntos 7, 8 y 9).
10. Resultado de la nota obtenida en la prueba de admisión de Lengua.
11. Resultado de la nota obtenida en la prueba de admisión de Historia.
12. Resultado de la nota obtenida en la prueba de admisión de Inglés.
13. Resultado de la nota obtenida en la prueba de admisión de Razonamiento abstracto.
14. Calificación con la que se concurre al ingreso en los estudios pretendidos dentro de la Universidad Pontificia Comillas.
15. Admitido o no en primera opción dentro de los estudios del Grado en Administración y Dirección de Empresas.

b) Información acreditativa de los requisitos de acceso o procedente de la estadística universitaria que los alumnos admitidos han de cumplimentar obligatoriamente al formalizar su primera matrícula al inicio del curso académico:

16. El alumno ha tenido un trabajo remunerado en el último año o no. Con posibles valores: 1, no realizó ningún trabajo o actividad remunerada; 2, trabajo a jornada completa (durante más de tres meses); 3, trabajo a jornada parcial (durante más de tres meses); 4, trabajo esporádico (durante menos de tres meses).
17. Forma de acceso para extranjeros (SI/NO).
18. Especialidad cursada dentro de los estudios realizados para acceder al sistema universitario, con posibles valores: 1, ciencias de la naturaleza y de la salud; 2, ciencias y/o tecnología; 3, humanidades y/o ciencias sociales;
19. Nivel de estudios del padre, con posibles valores: 1, Estudios primarios; 2, estudios secundarios; 3, estudios superiores; 4, sin estudios.
20. Nivel de estudios de la madre, con posibles valores iguales a los de la variable anterior.
21. Calificación, dentro de la fase general, de la PAU.

c) Información que da cuenta de los primeros resultados obtenidos en el transcurso de los estudios del alumno:

22. Superación o no en primera convocatoria la asignatura Fundamentos de Gestión Empresarial.

23. Superación o no en primera convocatoria la asignatura Marketing.
24. Superación o no en primera convocatoria la asignatura Matemáticas Empresariales I.
25. Superación o no en primera convocatoria la asignatura Idiomas I.
26. Número de asignaturas aprobadas en el primer cuatrimestre.

5. Resultados

A continuación se exponen los resultados obtenidos con el propósito de identificar alumnos con riesgo alto de cursar baja o presentar un bajo rendimiento académico en primer curso. Los resultados en este epígrafe cubren el objetivo del presente trabajo: estudiar si del análisis de la información de los solicitantes, basado en árboles de clasificación, es factible determinar un perfil del estudiante con riesgo de cursar baja en el momento de la admisión de aquél, del inicio del curso académico y tras la realización de los primeros exámenes, que sirva de ayuda en los procesos de tutorización de estudiantes en la universidad.

En los estudios de Grado en Administración y Dirección de Empresas, el porcentaje de alumnos que bien ha repetido primer curso o ha cursado baja al finalizar su primer año de estudios es del 12,9% (n=844). En las tablas 2 y 3 se presenta el resultado de los correspondientes contrastes (Gamma, o Phi y V de Cramer, según corresponda) para determinar el nivel de asociación entre variables cualitativas y la variable bajo rendimiento académico. Aquellas variables que presentan nivel de asociación son: el sexo de los estudiantes, la superación o no en primera convocatoria de las asignaturas del primer cuatrimestre, el número de éstas, la especialidad cursada en los estudios de acceso al Grado y el nivel de estudios del padre o tutor.

Tabla 2. Contrastes de nivel de asociación con variables ordinales y variable bajo rendimiento académico

Contraste de asociación, ordinal por ordinal	**Valor estadístico Gamma**	**Sig. aproximada**
Sexo (0, mujer; 1, hombre) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	,400	**,000**
Admitido en primera opción (0, No; 1, Si) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	-,124	,228
Acceso extranjeros (0, No; 1, Si) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	,133	,539
Aprueba asig. Fundamentos Gestión Empresarial 1ª convocatoria (0, No; 1, Si)* Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	-,854	**,000**
Aprueba asig. Marketing 1ª convocatoria (0, No; 1, Si) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	-,826	**,000**
Aprueba asig. Matemáticas Empresariales 1ª convocatoria (0, No; 1, Si) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	-,813	**,000**
Aprueba asig. Idioma I 1ª convocatoria (0, No; 1, Si) * Bajo Rendimiento Académico (0, pasa a segundo; 1, cursa baja o repite).	-,680	**,001**

Tabla 3. Contrastes de nivel de asociación con variables nominales y variable bajo rendimiento académico

Variables para contraste de asociación, ordinal por ordinal	**Valor estadísticos Phi y V de Cramer**	**Sig. aproximada**
Trabajó último año * Bajo Rendimiento Académico.	,090	,085
Especialidad Acceso * Bajo Rendimiento Académico (pasa a segundo; cursa baja o repite).	,102	**,033**
Nivel Estudios Madre * Bajo Rendimiento Académico (pasa a segundo; cursa baja o repite).	,040	,778

Nivel Estudios Padre* Bajo Rendimiento Académico (pasa a segundo; cursa baja o repite).	,116	**,026**
Número Asignaturas superadas en 1ª convocatoria * Bajo Rendimiento Académico (pasa a segundo; cursa baja o repite).	,514	**,000**

En la figura 2 se muestran los correspondientes gráficos de distribución de frecuencias entre la variable objetivo (bajo rendimiento académico) y aquellas variables que han presentado evidencias de existencia de nivel de asociación. Se destaca que el porcentaje de bajo rendimiento académico es superior en alumnos respecto de las alumnas, también entre quienes acceden tras haber cursado estudios vinculados con una especialidad de Ciencias Sociales y Humanidades frente a quienes lo hacen tras haber cursado una especialidad de Ciencias y/o Tecnología. Asimismo, se presentan porcentajes altos de aprobados en todas las asignaturas del primer cuatrimestre, salvo en la asignatura de Matemáticas Empresariales.

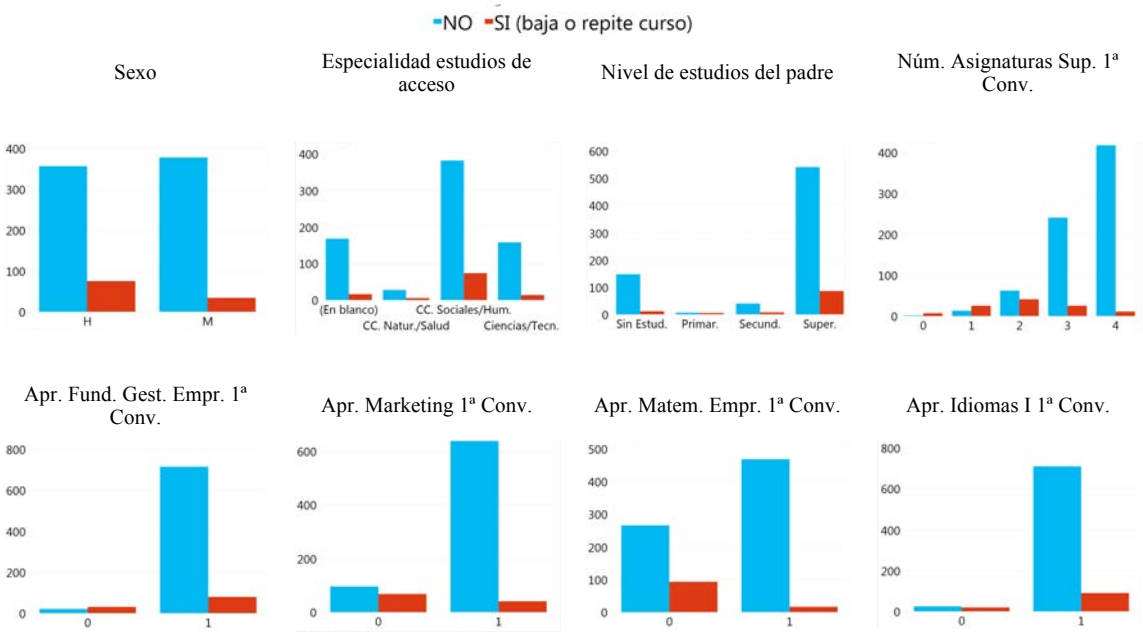

Figura 2. Gráficos de distribución de frecuencias de variables con nivel de asociación con la variable objetivo

En la figura 3 se puede observar como el perfil del alumno que ingresa en el Grado en Administración y Dirección de Empresas es el de un estudiante con nota media de expediente preuniversitario entre 7 y 8 (sobre diez). Se señala también que las calificaciones obtenidas en las pruebas de admisión relativas a Matemáticas son bajas, con una media de calificaciones que incluso se sitúan por debajo del aprobado. Este factor parece anticipar el elevado porcentaje de suspensos que se presenta posteriormente en la primera convocatoria de la asignatura de Matemáticas Empresariales.

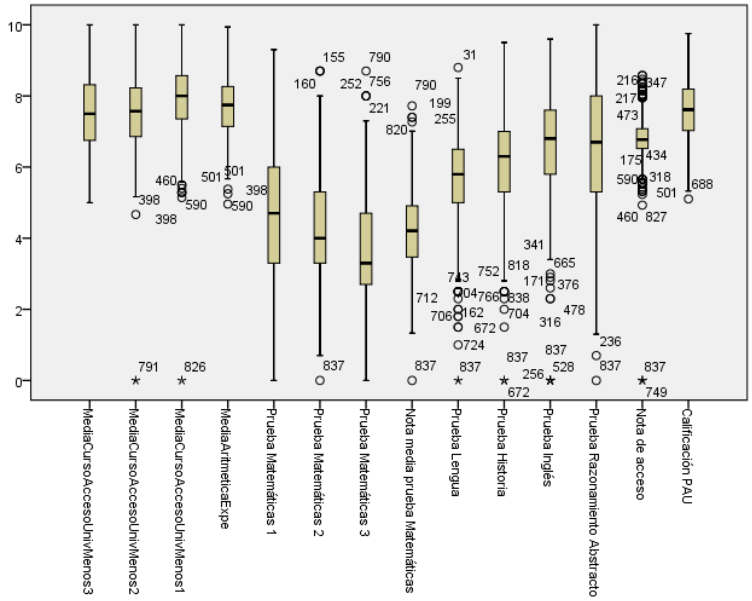

Figura 3. Gráficos de caja para variables de escala.

Respecto del porcentaje de aprobados y suspensos que se presenta en la primera convocatoria de las asignaturas del primer semestre, según se recogen en la tabla 4, la asignatura donde más suspensos hay en primera convocatoria es Matemáticas Empresariales I (42,2%), seguida de la asignatura de Marketing (19%). Por su parte, los bajos porcentajes de suspensos en las asignaturas de Fundamentos de Gestión Empresarial e Idioma extranjero I apuntan a ser variables con poca capacidad de segmentación. Asimismo, se observa que se produce un aumento importante en el porcentaje de suspensos de aquellos alumnos que terminan presentando bajo rendimiento académico (repiten o cursan baja) cuando superan menos de 2 asignaturas en la primera convocatoria de asignaturas del primer cuatrimestre.

Tabla 4. Porcentajes de suspensos y probabilidad de bajo rendimiento académico en el primer año de los estudiantes de nuevo ingreso del Grado en Administración y Dirección de Empresas por asignatura

		% Suspensos	% de Suspensos con bajo rendimiento
Asignatura del primer cuatrimestre.	Fundamentos de Gestión Empresarial	5,7%	58,3%
	Marketing	19%	40,3%
	Matemáticas Empresariales I	42,2%	24,9%
	Idioma extranjero I	4,8%	40,0%
Número de asignaturas superadas	0	0,7%	83,3%
	1	4,1%	64,7%
	2	12,1%	38,6%
	3	31,7%	9,5%
	4	51,4%	2,6%

Resultado 1ª convocatoria. Cursos 2009/2010 a 2014/2015.

Tal y como ya se ha indicado, el análisis de segmentación se ha replicado con información disponible en tres momentos: cuando el alumno es admitido, al inicio del curso académico y tras la realización de los primeros exámenes. El objeto de tal medida no es otro que servir de apoyo en los procesos de seguimiento y tutorización de estudiantes en la universidad, identificando lo antes posible qué características presentan los estudiantes que tendrían mayor probabilidad de tener un bajo rendimiento académico.

Realizados los análisis de segmentación con las variables procedentes del estudio exploratorio, se señala que en el momento de la admisión del estudiante y cuando comienza el curso académico no se produce segmentación alguna por parte de los algoritmos CART o QUEST. Sin embargo, no ocurre lo mismo cuando se realiza el análisis de segmentación incluyendo las variables de los primeros resultados oficiales obtenidos por los estudiantes de nuevo ingreso del Grado en Administración y Dirección de Empresas. En la figura 4 se muestra la forma que presenta el árbol de clasificación obtenido, apreciándose que el primer nivel de clasificación se segmenta en función de si se ha suspendido o no la asignatura de Marketing. Así, entre quienes suspenden dicha asignatura y también suspenden la asignatura de Matemáticas empresariales se tiene un 52,9% de frecuencia de ocurrencia de presentar bajo rendimiento académico. Es destacable el hecho de que la primera cortadura del árbol sea la asignatura de Marketing y no la de Matemáticas Empresariales I, a pesar de que esta última presentaba un porcentaje de suspensos superior.

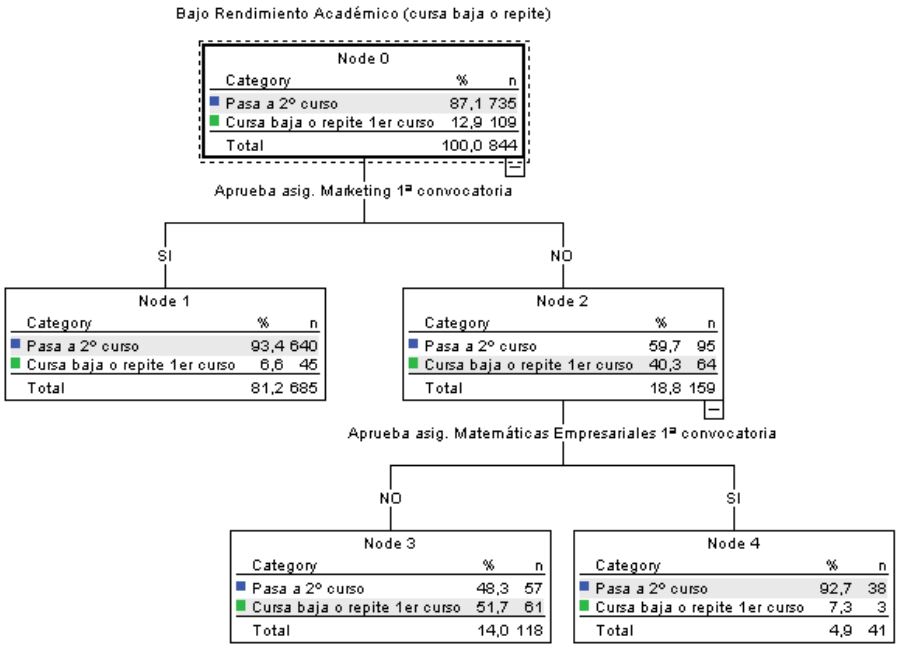

Figura 4. Segmentación del abandono al finalizar el primer curso en el Grado en Administración y Dirección de Empresas con variables disponibles al finalizar el primer cuatrimestre

El porcentaje de acierto sobre la categoría de mayor interés (bajo rendimiento académico en el primer año), tal y como señala la tabla 5, se sitúa en un 56%, que corresponde a los 61 alumnos del nodo 3 que han cursado baja o han repetido en su primer año, respecto del total de estudiantes que lo han hecho. En consecuencia, el modelo estima que aquellos estudiantes del Grado en Administración y Dirección de Empresas que no han superado las asignaturas de Marketing y Matemáticas Empresariales I han de ser clasificados como potenciales alumnos con bajo rendimiento académico.

En cuanto a la validez del modelo, el riesgo obtenido es bajo (0,181). No obstante, ha de tenerse en cuenta la baja frecuencia que se presenta en la categoría de cursar baja o repetir primer curso entre los

estudiantes del Grado en Administración y Dirección de Empresas. Este hecho reduce *per se* la probabilidad de clasificar incorrectamente individuos.

Tabla 5 Tabla de clasificación de acierto y riesgo del árbol obtenido, con información presente al finalizar el primer cuatrimestre. Alumnos de nuevo ingreso del Grado en Administración y Dirección de Empresas, cursos 2009/10- 2014/15

Real	Estimado		
	Año siguiente pasa a 2º curso	Cursa baja o Repite 1er curso	Porcentaje Acierto
Año siguiente pasa a 2º curso	678	57	92,2%
Cursa baja o Repite 1er curso	48	61	56,0%
Porcentaje Global	86,0%	14,0%	87,6%
Riesgo			
Estimado	Std. Error		
,181	,018		

Otros estudios, como el realizado por Ortiz (2015) entre los estudiantes del Grado en Ingeniería Electromecánica, sí han dado lugar a segmentaciones en los tres momentos analizados, aumentando la tasa de clasificación a medida que la información se acumulaba en cada momento considerado.

De la información obtenida en los estudios llevados a cabo sobre los alumnos de nuevo ingreso del Grado en Administración y Dirección de Empresas se pueden definir 3 niveles de riesgo de repetir primer curso o causar baja en el primer año de estudios. La composición de los niveles de riesgo, la probabilidad de cursar baja en cada uno de ellos, la frecuencia de ocurrencia dentro del colectivo de alumnos, así como las características que en cada caso presentan los estudiantes que agrupa cada nivel se presentan en la tabla 6.

Tabla 6. Clasificación de niveles de riesgo de presentar un bajo rendimiento académico en el primer año de estudios por los alumnos de nuevo ingreso del Grado en Administración y Dirección de Empresas

Nivel riesgo	Probabilidad repetir 1º /cursar baja	Frecuencia ocurrencia	Características de los estudiantes identificados
1	83,3%	0,7%	No superar ninguna asignatura en primera convocatoria del plan de estudios oficial correspondiente al primer cuatrimestre.
2	64,7%	4,1%	Superar una sola asignatura en primera convocatoria del plan de estudios oficial correspondiente al primer cuatrimestre.
3	51,7%	14,0%	Suspender en primera convocatoria la asignatura de Marketing y de Matemáticas Empresariales I.

Identificados los perfiles de riesgo, así como los alumnos pertenecientes a cada perfil se propone llevar a cabo una intervención por parte de los agentes oportunos de la universidad, por ejemplo a través de los correspondientes tutores de grupos, que traten de cambiar el destino pronosticado con el presente modelo.

6. Conclusiones

Con el propósito de determinar si del análisis de la información de los solicitantes es factible poder extraer un perfil del estudiante con riesgo de cursar baja o presentar un bajo rendimiento académico con margen suficiente para ayudar en los procesos de tutorización y seguimiento de los estudiantes en la universidad se ha tomado como muestra el conjunto de alumnos de nuevo ingreso en los estudios del Grado en Administración y Dirección de Empresas de la Universidad Pontificia Comillas (cursos 2009-2010 a 2014-2015).

Los hallazgos más relevantes del estudio llevado a cabo, que ha combinado la ejecución de un estudio exploratorio inicial con un posterior análisis de segmentación apoyado sobre los algoritmos CART y QUEST, son los siguientes:

- El porcentaje de bajo rendimiento académico es superior en hombres y en quienes han cursado estudios vinculados con una especialidad de Ciencias Sociales y/o Humanidades. También lo es entre quienes no superan alguna de las asignaturas del primer semestre, si bien hay diferencias en función de la asignatura de la que se trate. La influencia que tiene la especialidad cursada en los estudios que dan acceso al nivel universitario sobre el rendimiento académico son coherentes con, entre otros, Castellanos Val *et al*. (1998), Herrera *et al*. (1999) y Rúa y Kennedy (2003
- Se ha obtenido un 56% de éxito en la clasificación de los estudiantes que terminan presentando bajo rendimiento académico, en base a la información disponible al finalizar el primer cuatrimestre.
- No ha sido posible la segmentación de estudiantes en riesgo de bajo rendimiento académico con información presente en un momento anterior a la conclusión del primer cuatrimestre.
- Se han podido definir 3 niveles de riesgo de repetir primer curso o causar baja en el primer año de estudios.

En síntesis, en este trabajo se ha visto la utilidad que la técnica de los árboles de clasificación presenta para ayudar en la identificación temprana de aquellos alumnos con riesgo de presentar un bajo rendimiento académico -o abandonar sus estudios- en su primer año de estudios, para poder llevar a cabo sobre ellos una labor de tutorización y asesoramiento específico.

Los hallazgos obtenidos señalan la capacidad que determinadas variables, con las que cuentan las universidades, pueden presentar para alertar tempranamente de altos niveles de riesgo de bajo rendimiento académico. En este sentido, en el presente trabajo lo han sido haber superado las materias de Marketing y de Matemáticas Empresariales I en primera convocatoria, si bien, la capacidad de inferencia de estas dos variables ha de circunscribirse exclusivamente al estudio realizado.

Por su parte, las posibles razones para la obtención de tasas de acierto, menores respecto, por ejemplo a las encontradas entre los estudiantes del Grado en Ingeniería Electromecánica por Ortiz (2015), apuntarían al reducido porcentaje de alumnos en la categoría objetivo (109), al número de individuos considerados dentro del estudio (menor de 900), así como a la propia estructura del plan de estudios en su primer cuatrimestre. Las dos primeras constituyen una limitación de la técnica de los árboles de clasificación que ha de ser considerada.

Cabe señalar igualmente que la muestra tomada, restringida a alumnos de nuevo ingreso del Grado en Administración y Dirección de Empresas de la Universidad Pontificia Comillas no permite la generalización de los resultados obtenidos. Esta es, sin duda, una limitación importante que pretende ser minimizada en futuros trabajos. No obstante, dicha limitación no impide que, aunque las diferencias entre planes de estudios hagan imposible actualmente la consideración de un modelo único universal para un título, sí sea factible, en cambio, que cada universidad replique la metodología presentada en este trabajo, según sus particularidades, posibilidades y contexto. Con los resultados obtenidos podrá determinar de forma particular cuando, específicamente, cada universidad podría realizar una identificación temprana de

sus alumnos en situación de riesgo de bajo rendimiento que sirva de ayuda a los programas de tutorización, asesoramiento y *mentoring*.

Asimismo, si bien la iniciativa planteada no asegura *per se* que las tasas de rendimiento mejoren, sí aumenta la probabilidad de que ello ocurra (Herzog, 2006). Por esta razón, además de la importancia que adquiere la correcta elección de los programas de tutorización, asesoramiento y mentoring que fuesen a ser desarrollados, se sugiere como futura línea de investigación profundizar en el impacto real que aquéllos tienen sobre las tasas de rendimiento, así como la vinculación entre la iniciativa planteada y la influencia del denominado capital social por parte del resto de estudiantes.

Por último, surge también el interrogante acerca de si la iniciativa planteada cuenta con impedimentos para llevarse a cabo por parte de sus destinatarios. Es decir, previamente a la ejecución de la iniciativa habría de poder responderse a las siguientes preguntas: ¿los estudiantes desean saber si se encuentran en situación de riesgo de tener bajo rendimiento académico con el objeto de aplicar sobre ellos medidas específicas de ayuda y seguimiento?, y, en tal caso, ¿cuál sería la credibilidad que les merece la iniciativa? La obtención de un nivel de respuesta negativo en una o ambas cuestiones podría limitar las posibilidades de aplicabilidad en la universidad. Ambas cuestiones quedan vinculadas con distintos aspectos asociados a la gestión de la calidad de las universidades que pueden ser objeto de una futura línea de investigación.

Agradecimientos

A la Facultad de Ciencias Económicas y Empresariales de la Universidad Pontificia Comillas por la colaboración para poder llevar a cabo esta investigación en su centro.

Referencias Bibliográficas

1. L. Cabrera, J. T. Bethencourt, P. Á. Pérez y M. G. Alfonso, (2006). El problema del abandono de los estudios universitarios. *Relieve*, Recuperado a partir de http://www.uv.es/relieve/v12n2/RELIEVEv12n2.
2. A. Cernuda del Río, S. Hevia Vázquez, M. C. Suárez Torrente y D. Gayo Avello, (2007), Un estudio sobre el absentismo y el abandono en asignaturas de programación, in *XIII Jornadas de Enseñanza Universitaria de Informática*.
3. Z. L. Berge y Y. P. Huang, (2004). A Model for Sustainable Student Retention: A Holistic Perspective on the Student Dropout Problem with Special Attention to e-Learning. *Deosnews*, *13*(5).
4. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis y V. Loumos, (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques, *Computers & Education*, *53*(3), pp. 950–965.
5. P. O'keeffe, (2013). A Sense of Belonging: Improving Student Retention, *College Student Journal*, *47*(4), 605-613.
6. M. Yorke, (2004). *Leaving early: Undergraduate non-completion in higher education* (Falmer Press, Londres).
7. V. Tinto, (1975). Dropout from higher education: A theoretical synthesis of recent research, *Review of educational research, 45*(1), pp. 89–125.
8. Yasmin, (2013). Application of the classification tree model in predicting learner dropout behaviour in open and distance learning, *Distance Education*, *34*(2), pp. 218-231.
9. T. Bartual y M. C.Poblet, (2009). Determinantes del rendimiento académico en estudiantes universitarios de primer año de Economía, *Revista de Formación e Innovación Educativa Universitaria (REFIEDU)*, *2*(3), pp. 305–314.
10. E. Corominas, (2001). La transición de los estudios universitarios: Abandono o cambio en el primer año de Universidad, *Revista de investigación educativa, RIE*, *19*(1), pp. 127–152.
11. P. A. Murtaugh, L. D. Burns y J. Schuster, (1999). Predicting the retention of university students, *Research in Higher Education*, *40*(3), pp. 355-371.

12. A. J. Adam y G. H. Gaither, (2005). Retention in higher education: A selective resource guide, *New Directions for Institutional Research,* 125, pp. 107-122.
13. L. W. Boyles, (2000). Exploration of a retention model for community college students, Ph D. thesis (University of North Carolina, Estados Unidos de América).
14. L. Hagedorn (2005), How to define retention, *College student retention formula for student success*, Plymouth: Rowman & Littlefield Publishers, pp. 90–105.
15. S. Herzog, (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression, *New Directions for Institutional Research*, 131, pp. 17-33.
16. J. Mohammadi (1994), *Exploring Retention and Attrition in a Two-Year Public Community College*, (Patrick Henry Community College), Recuperado a partir de http://files.eric.ed.gov/fulltext/ED382257.pdf
17. S. Rodríguez, E. Fita y M. Torrado, (2004). El rendimiento académico en la transición secundaria-universidad, *Revista de educación*, 334, pp. 391–414.
18. Texas State Higher Education Coordinating Board, (2004). *The Art of Student Retention: A Handbook for Practitioners and Administrators* (Texas Higher Education Coordinating Board).
19. V. Tinto, (2006). Research and practice of student retention: what next?, *Journal of College Student Retention: Research, Theory and Practice*, *8*(1), pp. 1–19.
20. V. Tinto (2007), *Taking student retention seriously*, (Syracuse University, Estados Unidos de América). Recuperado a partir de http://www.umesgolf.com/assets/0/232/3812/4104/4110/bd28b4ae-e1cc-4575-9b37-535d2d2be5f1.pdf.
21. C. Vivian, (2005). Advising the At-Risk College Student, The *Educational Forum, 69*(4), pp. 336-351.
22. L. Wild y L. Ebbers, (2002). Rethinking student retention in community colleges, *Community College Journal of Research & Practice*, *26*(6), pp. 503–519.
23. V. Tinto, (1998). Stages of student departure: Reflections on the longitudinal character of student leaving, *The Journal of Higher Education, 59*(4), pp. 438–455.
24. W. S. Swail, (2004). The art of student retention: A handbook for practitioners and administrators, *20th Annual Recruitment and Retention Conference*. Recuperado a partir de http://files.eric.ed.gov/fulltext/ED485498.pdf.
25. R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri y M. J. Ramírez-Quintana, (2007). Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos, in *XIII Jornadas de Enseñanza universitaria de la Informática*.
26. M. García y M. J. San Segundo, (2001). El rendimiento académico en el primer curso universitario, in *X Jornadas de Economía de la Educación*.
27. M. A. Guisande, A. P. Soares, L. S. Almeida y A. Diniz, A., (2006). Construcción y validación de un modelo multidimensional de ajuste de los jóvenes al contexto universitario, *Psicothema*, *18*(2), pp. 249-255.
28. M. N. Rodríguez y M. T. Coello, (2008). Prediction of university students' academic achievement by linear and logistic models, *The Spanish journal of psychology*, *11*(1)., 275–288.
29. B. Shulruf,J. Hattie y S. Tumen, (2008). The predictability of enrolment and first-year university results from secondary school performance: the New Zealand National Certificate of Educational Achievement, *Studies in Higher Education*, *33*(6), 685-698.
30. E. Carrión, (2002). Validación de características al ingreso como predictores del rendimiento académico en la carrera de medicina, *Educación Médica Superior*, *16*(1), pp. 1–2.
31. M. A. Goberna y M. A. López, (1987). La predicción del rendimiento como criterio para el ingreso en la Universidad, *Revista de educación*, 283, 235–248.
32. R. Peña y I. Sánchez, (2005). Análisis estadístico del rendimiento académico de una asignatura con relación a asignaturas anteriores, in *XI Jornadas de Enseñanza Universitaria de Informática*.
33. A. Rúa y L. Kennedy, (2003). Influencia de la formación preuniversitaria en el rendimiento del 1º curso de Ciencias Empresariales Europeas (E4), in *X Jornadas de ASEPUMA. Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa*.. Recuperado a partir de http://www.uv.es/asepuma/X/comunica.htm.

34. L. A. Díaz y C. R. Toloza, (2007). Los indicadores de selección para el ingreso a la universidad y su valor para estimar el rendimiento académico en el primer semestre, *CIMEL*, *12*(2), pp. 59–65.
35. J. R. Betts y D. Morell, (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects, *Journal of Human Resources,* 34, pp. 268-293.
36. A. Cortés, (2008). El proceso de admisión como predictor del rendimiento académico en la educación superior, *Universitas Psychologica, 7*(1), pp. 197-213.
37. M. V. García, A. Jiménez y J. M. Alvarado, (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística, *Psicothema*, *12*(2), pp. 248–252.
38. R. Gimeno, R. Redondo y A. Rúa, (2003). ¿Es redundante la prueba de selectividad?, in *XI Jornadas de ASEPUMA. Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa*, Recuperado a partir de https://doaj.org/toc/1575-605X/Actas_11.
39. O. D. Marcenaro y M. L Navarro, (2007). El éxito en la universidad: Una aproximación cuantílica, *Revista de Economía Aplicada*, *15*(44), pp. 5-39.
40. A. Vélez van Meerbeke y C. N. Roa González, (2005). Factores asociados al rendimiento académico en estudiantes de medicina, *Educación Médica*, *8*(2), pp. 24–32.
41. A. Cuxart Jardí, M. Martí Recober y F. Ferrer Juliá, (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las Pruebas de Aptitud de Acceso a la Universidad (PAAU), *Revista de Educación*, 314, pp. 63-88.
42. T. Escudero Escorza, (1987). Buscando una mejor selección de universitarios, *Revista de Educación*, 283, pp. 249-283.
43. L. Castellanos Val, M. C. González Veiga, M. A. González de Sela Isabel y M. Manzano Pérez, (Santiago de Compostela, 1998, septiembre) Las matemáticas empresariales: estudio de los factores determinantes del rendimiento académico, in *VI Jornadas de ASEPUMA. Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa*, Recuperado a partir de http://www.uv.es/asepuma/VI/17.PDF.
44. T. Escudero Escorza, (1984). Condicionantes y capacidad predictiva de la selectividad universitaria, *Revista de Educación*, 273, pp. 139-164.
45. M. E. Herrera, S. Nieto, M. J. Rodríguez y M.C. Sánchez, (1999). Factores implicados en el rendimiento académico de los alumnos, Universidad de Salamanca, *Revista de Investigación Educativa*, 17(2), pp. 413–421.
46. J. Mafokozi, L. O. Martín-Varés, C. González y A. de la Orden Hoz, (2001). Modelos de investigación del bajo rendimiento, *Revista complutense de educación*, *12*(1), pp. 159–178.
47. J. L. García Llamas, El análisis discriminante y su utilización en la predicción del rendimiento académico, *Revista de Educación*, *280* (1986), pp. 229-252.
48. A. Rúa, R. Redondo y R. Gimeno, (2003). Tipología de los alumnos que entran en la universidad española, in *XI Jornadas de ASEPUMA, Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa*, Recuperado a partir de http://www.uv.es/asepuma/XI/48.pdf.
49. J. B. Gallestey, M. R. Gil y J. B. Guerra, (2004). Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico, *Revista Cubana de Educación Médica Superior*, *18*(3), pp. 1-11.
50. A. Rúa y C. González*,* (2004). Predicción del rendimiento académico final a partir de pruebas previas en asignaturas cuantitativas, in *XII Jornadas de ASEPUMA*. *Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa*, Recuperado a partir de http://www.um.es/asepuma04/.
51. T. Bartual y M. C. Poblet, (2009). Determinantes del rendimiento académico en estudiantes universitarios de primer año de Economía, *Revista de Formación e Innovación Educativa Universitaria (REFIEDU)*, *2*(3), pp. 305–314.
52. L. A. Díaz y C. R. Toloza (2007), Los indicadores de selección para el ingreso a la universidad y su valor para estimar el rendimiento académico en el primer semestre, *CIMEL*, *12*(2) (2007), pp. 59–65.
53. S. J. Press y S. Wilson, (1979). Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association, 73*(364), pp. 699-705.

54. M. E. Mercado, (2012). Las aplicaciones del análisis de segmentación: El procedimiento Chaid, *Empiria, Revista de metodología de ciencias sociales*, 1 (2012), pp. 13–49.
55. B. K. Baradwaj y S. Pal, (2012). Mining educational data to analyze students' performance, *International Journal of Advanced Computer Science and Applications 2* (6), 63-69, Recuperado a partir de http://arxiv.org/abs/1201.3417.
56. V. Berlanga, M. J. Rubio Hurtado y R. Vilà Baños, (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE, Revista d'Innovació i Recerca en Educació, 6*(1), pp. 65-79, Recuperado a partir de http://diposit.ub.edu/dspace/handle/2445/43762.
57. L. Jing, (2012). Data Mining and Its Applications in Higher Education. *New Directions for Institutional Research*, 113, pp. 17-36.
58. C. Pérez y D. Santín, (2007). *Minería de datos. Técnicas y herramientas* (Thompson, Madrid).
59. J. M. Rojo (2006). *Árboles de Clasificación y Regresión* (Instituto de Economía y Geografía. Consejo Superior de Investigaciones Científicas, Madrid.
60. J. F. Superby, J. P. Vandamme y N. Meskens, (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods, *ITS'06 Proceedings of the 8th international conference on Intelligent Tutoring Systems,* pp. 37–44.
61. J. P. Vandamme, N. Meskens y J. F. Superby, (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics, 15*(4), pp. 405-419.
62. F. Angúlo y E. Sergio, (2012) Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios PhD. The sis (Universidad de Chile, Chile). Recuperado a partir de http://www.tesis.uchile.cl/handle/2250/111188.
63. J. P. Campbell, P. B. DeBlois y D. G. Oblinger, (2007). Academic analytics: A new tool for a new era. *Educause Review*, *42*(4), pp. 40-50.
64. G. W Dekker, M. Pechenizkiy y J. P. Vleeshouwers, (2009). *Predicting Students Drop Out: A Case Study*. *Educational Data Mining*, pp. 41-50.
65. J. P. Gallestey, M. R. Gil y J. B. Guerra, (2004). Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico. *Revista Cubana de Educación Médica Superior*, *18*(3), 1-11.
66. S. Herzog, (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression, *New Directions for Institutional Research*, 131, pp. 17-33.
67. S. A. Kumar y M. N. Vijayalakshmi, (2011). Efficiency of decision trees in predicting student's academic performance, in First International Conference on Computer Science, Engineering and Applications.
68. M. E. Mercado (2017), *El análisis de segmentación: técnica y aplicaciones de los árboles de clasificación*. (Madrid, Centro de Investigaciones Sociológicas).
69. N. T. Nghe, P. Janecek y P. Haddawy, (2007). A comparative analysis of techniques for predicting academic performance, *Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports*.
70. E. Porcel, G. N. Dapozo y M. V. López, (San Juan, Argentina, 2009). Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios, in XI Workshop de Investigadores en Ciencias de la Computación. Recuperado a partir de http://hdl.handle.net/10915/19846.
71. M. M. N. Quadri y D. N. V. Kalyankar, (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology*, *10*(2), pp. 2-5, Recuperado a partir de http://computerresearch.org/stpr/index.php/gjcst/article/view/128.
72. M. Ramaswami y R. Bhaskaran, (2010). A CHAID Based Performance Prediction Model in Educational Data Mining, *International Journal of Computer Sciences Issues, 7*(1), pp. 10-18.
73. G. De'ath y K. E. Fabricius, (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), pp. 3178–3192.
74. A. P. Goicoechea, (2002). *Imputación basada en árboles de clasificación*. Eustat, Instituto Vasco de Estadística. Recuperado a partir de http://www.eustat.es/document/datos/ct_04_c.pdf.

75. L Breiman, J. Friedman, C. J. Stone y R. A. Olshen, (1984). *Classification and regression trees*, (CRC press, Londres).
76. W. Y. Loh y Y. S. Shih, Split selection methods for classification trees, *Statistica sinica*, *7*(4) (1997), pp. 815–840.
77. R. Gnanadesikan, (1977). *Methods for statistical data analysis of multivariate observations* (Wiley, Vol. 321, Nueva York).
78. J. A. Hartigan y M. A. Wong, (1979). Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*(1), pp. 100–108.
79. J. M. Ortiz Lozano, (2015). Gestión de la calidad en el ámbito universitario. Una aproximación a la gestión de procesos asociados al ingreso de alumnos según el modelo EFQM, PhD. thesis (Universidad Pontificia Comillas de Madrid).