# Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos

R. Alcover[1], J. Benlloch[2], P. Blesa[3], M. A. Calduch[1], M. Celma[3], C. Ferri[3], J. Hernández-Orallo[3], L. Iniesta[5], J. Más[4], M. J. Ramírez-Quintana[3], A. Robles[2], J. M. Valiente[2], M. J. Vicent[3], L. R. Zúnica[1]

[1]Dpto. de Estadística e I.O. Aplicadas y Calidad, [2]Dpto. de Informática de Sistemas y Computadores, [3]Dpto. de Sistemas Informáticos y Computación, [4]Dpto. de Física Aplicada, [5]Becaria PACE

Universidad Politécnica de Valencia

Camino de Vera s/n, 46022 Valencia

jmas@fis.upv.es

## Resumen

En este trabajo presentamos un análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia (UPV) a lo largo de tres cursos, aunque también se ha trabajado con las titulaciones de Ingeniería Técnica en Informática de Gestión y de Ingeniería Informática.

Este análisis relaciona el rendimiento con las características socioeconómicas y académicas de los alumnos, que se obtienen en el momento de su matrícula, y que se recogen en la base de datos de la universidad. Hemos definido un indicador del rendimiento para cada alumno, teniendo en cuenta las calificaciones obtenidas y las convocatorias utilizadas.

Para el estudio utilizamos técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, etc… Esto proporciona una herramienta importante para la acción tutorial, que puede apoyarse en las predicciones de los modelos que se obtienen para encauzar sus recomendaciones y encuadrar las expectativas y el esfuerzo necesario para cada alumno, lógicamente dentro de la cautela habitual a la hora de tratar modelos inferidos a partir de datos.

## 1. Motivación

Desde hace algún tiempo se le está dedicando una creciente atención al rendimiento académico universitario. Esta mayor atención viene determinada por factores tanto de índole económica como política y social; la permanencia de los estudiantes en la universidad durante prolongados períodos de tiempo es un aspecto controvertido y causante, en parte, de la demanda cada vez mayor de las titulaciones de primer ciclo frente a las de segundo ciclo [6].

Por otra parte, el proceso de convergencia europea de la educación superior está obligando a las universidades al diseño de estrategias de adaptación debido a la profunda reforma que este proyecto implica. Los procesos de evaluación y acreditación de títulos puestos en marcha en los últimos años para llevar a término esta adaptación se sustentan en la construcción de indicadores, de muy variada índole, que permiten descubrir las fortalezas y debilidades de las enseñanzas universitarias actuales. Como se comenta en [9], "reflexionar sobre todos los elementos que la evaluación del rendimiento del alumnado proporciona se convierte en un mecanismo claro para la mejora de la calidad del proceso educativo". En este sentido, es interesante analizar el rendimiento académico de los estudiantes para tomar medidas oportunas tanto de forma individual como global. Asimismo, según [5], en la propuesta de los futuros títulos de Grado y Máster se deberá incluir una estimación de los valores de ciertos indicadores de resultados, como las tasas de graduación, abandono y eficiencia. En este aspecto, el Ministerio de Educación y Ciencia ya recoge una lista de indicadores [7] para estos estudios. Más aún, la viabilidad académica de los títulos (que estos puedan ser completados en el tiempo previsto por la mayoría de los estudiantes

a tiempo completo) será un criterio clave para la aprobación de los mismos, así como para la evaluación de las universidades y sus centros.

Algunos factores podrían, en gran medida, explicar el éxito o fracaso de un estudiante, como sus características socioeconómicas, edad, estudios previos, entorno al inicio de sus estudios, actividad, o no, laboral durante los estudios, características organizativas y docentes de los centros, planes de estudios, métodos evaluativos, etc... Conocidos estos factores, tanto la universidad como los estamentos responsables de los estudios preuniversitarios, podrían estudiar acciones que mejoraran el rendimiento de colectivos específicos, como ya ocurre en el llamado Programa de Acción Tutorial Universitario (PATU). Este programa de tutorías, adoptado institucionalmente por nuestra universidad desde hace más de cinco cursos, pretende favorecer tanto la integración del alumno como su formación integral. Es en esta línea en la que se enmarca nuestro trabajo, para el cual hemos utilizado la investigación directa en las bases de datos que recogen información tanto sobre las características e historial del alumnado, como de la universidad, mediante técnicas de minería de datos [4]. A pesar de que existen innumerables trabajos descriptivos sobre rendimiento académico en universidades, no tenemos conocimiento de que las técnicas de minería de datos se hayan aplicado de forma exhaustiva a este tipo de estudios. Sí que aparece algún trabajo en el que se utiliza la técnica de *"clustering"* [1] o de regresión lineal mediante procedimientos tradicionales [3].

## 2. Metodología

La metodología que se ha seguido para la obtención del rendimiento se puede resumir en las siguientes etapas:

1. Establecer el/los objetivo/s del estudio.
2. Definir la población y la muestra de estudiantes implicada en el estudio.
3. Obtención de la vista minable, a partir de la información contenida en la base de datos de la universidad.
4. Elección del tipo de análisis de datos requerido.
5. Generación y validación de los modelos.
6. Interpretación de los resultados.

Seguidamente se presentan las tareas llevadas a cabo en cada una de estas etapas.

### 2.1. Objetivo del estudio

Nuestro estudio, a nivel global, pretende aplicar técnicas de minería de datos para analizar la influencia de los parámetros (socioeconómicos, características personales, nota de entrada…) más relevantes sobre el rendimiento académico de un alumno de primer curso en las titulaciones de informática de la UPV, de forma que nos permita predecir este rendimiento disponiendo únicamente de la información aportada por el alumno en el momento de su matrícula. Estas titulaciones corresponden a Ingeniería Informática (II) impartida por la Facultad de Informática (FI), Ingeniería Técnica en Informática de Gestión (ITIG) e Ingeniería Técnica en Informática de Sistemas (ITIS), estas últimas impartidas por la Escuela Técnica Superior de Informática Aplicada (ETSIAp). En este trabajo se presentan únicamente resultados correspondientes a ITIS.

La adopción de políticas encaminadas a corregir situaciones de fracaso académico a partir de las conclusiones de este trabajo no es un objetivo del estudio. Estas políticas deben ser acometidas bien por los centros, bien por los estamentos responsables.

### 2.2. Población

La población objeto de nuestro estudio está constituida por todos los alumnos de nuevo ingreso en cualquiera de las tres titulaciones de informática de la UPV antes mencionadas. Con el fin de trabajar con una muestra representativa de la población, se ha considerado a los alumnos de nuevo ingreso matriculados en primero de alguno de los títulos de informática durante los cursos 01-02, 02-03 y 03-04, esto es, desde el último cambio del plan de estudios. Así, el estudio se ha realizado sobre 569 alumnos de II, 646 alumnos de ITIG y 572 alumnos de ITIS.

### 2.3. Obtención de la vista minable

Como ya se ha comentado, la base de datos de una universidad incorpora, curso tras curso, una gran cantidad de información relativa al alumnado, procedencia, matrículas, calificaciones, planes de estudios,… Por lo tanto, esta base de datos

contiene información sobre diferentes factores, *a priori*, potencialmente relevantes para el rendimiento de un alumno. Desde los dos centros implicados en nuestro estudio se dispone de una vista parcial de la base de datos de la UPV que contiene la información de la FI y de la ETSIAp.

Un aspecto importante es la despersonalización de los datos puesto que, de acuerdo a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, se debe proceder tratando los datos personales de modo que la información que se obtenga no pueda asociarse a persona identificada o identificable. El proceso de despersonalización ha consistido en eliminar o sustituir cualquier información que pueda identificar directa o indirectamente al alumno (como el DNI, nombre y apellidos, domicilio, etc…).

Con el fin de crear un almacén de datos y un entorno que facilitara la obtención de datos para realizar el estudio, se decidió integrar los mismos en *Oracle*. De esta forma se ha podido utilizar como herramienta OLAP el *Oracle Discoverer*. Con ella se han extraído las vistas minables.

Una vista minable puede definirse como una colección de individuos sobre los cuales queremos realizar un determinado estudio, con todas sus características (atributos), que tiene como finalidad poder aplicar el proceso de la minería de datos sobre ella para poder extraer conocimiento útil. Así, hemos creado una vista minable por titulación. Cada una de ellas contiene las notas y datos personales de los alumnos de la muestra seleccionada. Para generarla, se ha utilizado el generador de informes del *Discoverer*, seleccionando, de entre todos los atributos disponibles, aquellos que se consideraron, a priori, que podrían tener mayor influencia en el rendimiento académico, filtrando el resto. Dichos atributos son:

- *Ocupacio P*: Ocupación del padre.
- *Ocupacio M*: Ocupación de la madre.
- *Ocupacio A*: Ocupación del alumno.
- *Ing Nota*: Nota con la que el alumno aprueba estudios de acceso.
- *Ing Est*: Estudios con los que accede a la titulación.

Seguidamente, se procedió a la agrupación de valores de algunos atributos por su elevado número de alternativas, con el fin de reducirlas y hacer más fácilmente interpretables los resultados obtenidos. Tales atributos son:

- *D_Altr Estud*: Otros estudios universitarios del alumno al ingresar en la titulación.
- *D_Estudis P*: estudios del padre.
- *D_Estudis M*: estudios de la madre.
- *Dpaises*: Derivado del país de nacimiento del alumno, agrupando por zonas geográficas.
- *Residencia Alumno*: Derivado de la provincia y el código postal donde reside el alumno durante el curso.
- *Residencia Familia Alumno*: Derivado de la provincia y el código postal donde reside la familia del alumno durante el curso.
- *Edad Ingreso*: Atributo derivado calculado como la diferencia entre el año de ingreso del alumno y año de nacimiento.

Finalmente, se especificó el tipo de cada atributo como nominal (o categórico) o numérico, siendo todos nominales excepto la nota de acceso a los estudios y la edad del alumno.

Además de todos estos datos personales del alumno, para completar la vista minable se incorporó una columna con el *Rendimiento* de cada alumno, *R*, calculado mediante la siguiente fórmula:

$$R = \frac{\sum_j 0,8^{c-1} \cdot Calif_j \cdot C_j \cdot 10}{\sum_j C_j}$$

donde $c$ es la convocatoria en que el alumno supera la asignatura (1 en el caso de ordinaria y 2 en el caso de extraordinaria). De esta forma penalizamos un 20% el rendimiento de los aprobados en segunda convocatoria; no caben más opciones en este primer curso. $Calif_j$ es la calificación numérica que obtuvo el alumno en la asignatura $j$ cuando aprobó dicha asignatura, tomando un 0 si el alumno no se presentó. En el caso de que el alumno suspendiera dicha asignatura, se toma la nota de la última convocatoria presentada. $C_j$ es el número de créditos con que figura la asignatura $j$ en el plan de estudios. Obsérvese que el rendimiento de un alumno así definido es un atributo que toma valores entre 0 y 100 [10].

### 2.4. Elección del tipo de análisis de datos requerido

Para lograr los objetivos marcados en nuestro trabajo (predecir el rendimiento de un alumno en su año de ingreso usando únicamente datos anteriores a su entrada en la universidad), los modelos de minería de datos que hemos elaborado son de tipo predictivo.

Hay que tener en cuenta que en minería de datos las técnicas estadísticas tradicionales pueden completarse con técnicas de inteligencia artificial. En este sentido, métodos adaptativos como los algoritmos genéticos y las redes neuronales permiten realizar predicciones muy acertadas, sobre todo en casos de gran complejidad y con relaciones internas no lineales, pero tienen el problema de ser menos legibles, por lo que no las hemos utilizado en nuestro trabajo.

De entre las técnicas de minería de datos existentes, hemos utilizado dos de ellas para generar los modelos predictivos del rendimiento: los árboles de decisión y la regresión multivariante.

- Los árboles de decisión son una serie de decisiones o condiciones organizadas de forma jerárquica, a modo de árbol. Son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclan datos categóricos y numéricos. Básicamente, un árbol de decisión es un árbol donde cada nodo representa una condición o test sobre algún atributo y cada rama que parte de ese nodo corresponde a un posible valor para ese atributo. Finalmente, las hojas representan el valor de la variable predicha. Esta técnica se usa en tareas de clasificación, agrupamiento y regresión. Los árboles de decisión que se usan para predecir variables categóricas se llaman árboles de clasificación, mientras que los árboles de decisión que se utilizan para predecir variables continuas se llaman árboles de regresión.
- La regresión multivariante es un método estadístico clásico que permite establecer una relación matemática entre un conjunto de variables independientes $X_1, X_2, ..., X_k$ y una variable dependiente $Y$. Se utiliza fundamentalmente en estudios en los que no se puede controlar por diseño los valores de las variables independientes. Los objetivos de un modelo de regresión pueden ser dos: obtener una ecuación que nos permita "predecir" el valor de $Y$ una vez conocidos los valores de las variables independientes, y cuantificar la relación entre las variables independientes y la dependiente con el fin de conocer o explicar mejor la relación. Se trata en este caso de modelos explicativos.

### 2.5. Generación y validación de los modelos

Para la generación de los modelos se ha utilizado la herramienta *SPSS Clementine v.9.0* [8]. En concreto, de los árboles de decisión que incorpora el Clementine, hemos utilizado para regresión el *árbol C&R*, que es un tipo de algoritmo de aprendizaje de árboles que se basa en el algoritmo CART de Leo Breiman et al. [2]. Este árbol realiza particiones binarias con el objetivo que la media de cada rama sea diferente y, por tanto, discrimine con la suficiente precisión en un número de particiones razonable como para poder asignar a cada hoja un valor cercano a la media de los elementos que caen en ella. Asimismo, también hemos aplicado el método *regression* del Clementine, que implementa una regresión lineal.

Una vez generados los modelos es necesario evaluarlos para determinar su calidad y utilidad. Dependiendo de la tarea de minería de datos existen diversos criterios que pueden usarse para evaluar los modelos, como, por ejemplo, la precisión predictiva (porcentaje de aciertos) generalmente utilizada en el contexto de la clasificación. Si la tarea es de regresión, tal y como la que nos ocupa, una forma de evaluar un modelo es mediante la raíz cuadrada positiva del error cuadrático medio (RECM), calculada como

$$RECM = \sqrt{ECM} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - R_i\right)^2}{n}}$$

es decir, la diferencia entre el valor que para el rendimiento predice el modelo ($y_i$) y el valor real ($R_i$) que podemos calcular con la fórmula establecida anteriormente, siendo $n$ el tamaño del conjunto de datos. Así, el parámetro obtenido viene expresado en las mismas unidades que los datos originales.

Si la cantidad de datos lo permite, la forma de entrenar y validar un modelo consiste en partir aleatoriamente los datos en dos subconjuntos disjuntos: el de los datos de entrenamiento

(*training set*), con el que se genera el modelo, y el de prueba o test (*test set*), con el que se evalúa el modelo. En ese caso, el RECM se calcula únicamente para los datos del conjunto de test.

En nuestro estudio, se ha realizado la siguiente partición de los datos: un 77% de los mismos se ha utilizado para el entrenamiento del modelo, y el 23% restante para validarlo.

A continuación mostramos, a modo de ejemplo, los modelos obtenidos para la vista minable de ITIS, así como sus errores.

La Figura 1 muestra el árbol de decisión generado por la herramienta.

```
D_Altr Estud in [ 1 ] => 58,0 (16)
D_Altr Estud in [ 2 ]  (424)
    Ing Nota <= 7,390  (339)
        Edad Ingreso <= 18,500 (160)
            Ing Nota <= 6,420 (82)
                D_Estudis P in [ 2 ] => 20,482 (26)
                D_Estudis P in [ 1 3 ] => 28,907 (56)
            Ing Nota > 6,420 (78)
                Ing Est in [ 2 ] => 25,854 (24)
                Ing Est in [ 10 ] => 40,863 (54)
        Edad Ingreso > 18,500 => 23,939 (179)
    Ing Nota > 7,390 (85)
        Ocupacio A in [ 1 ] => 28,229 (18)
        Ocupacio A in [ 2 3 ] (67)
            Edad Ingreso <= 19,500 => 50,43 (20)
            Edad Ingreso > 19,500 (47)
                D_Estudis_ M in [1 2] => 33,7 (36)
                D_Estudis_ M in [3] => 52,94 (11)
```

Figura 1. Árbol C&R para la titulación de ITIS

La notación que se sigue es: cada línea corresponde a un nodo del árbol, y contiene el nombre del atributo usado en ese nodo así como su/s valor/es entre corchetes, si es un atributo nominal, o bien una expresión de la forma ≤ V ó >V, si es un atributo numérico, siendo V un valor comprendido entre los valores mínimo y máximo de ese atributo en los ejemplos en ese nodo; además, si se trata de una hoja, se añade el símbolo "=>" tras el que se indica el valor predicho (el rendimiento, en nuestro caso); asimismo, el número que aparece entre paréntesis indica el número de instancias del conjunto de entrenamiento en ese nodo del árbol. Por ejemplo, en la Figura 1, el primer atributo seleccionado es *D_Altr Estud.* El primer hijo, que corresponde a aquellas instancias que tienen un valor de 1 (estudios universitarios) para este atributo (en concreto 16), es una hoja y el valor medio predicho para el rendimiento en este caso es 58. El otro nodo hijo es el que corresponde a las 424 instancias cuyo atributo *D_Altr Estud* tiene un valor de 2 (otros estudios). Este último nodo continúa expandiéndose ya que no es una hoja. A partir de los errores calculados por el Clementine para este modelo con respecto al conjunto de test, hemos obtenido la raíz cuadrada del error cuadrático medio (que no está incluido en la herramienta) que resultó ser 17,95.

El modelo de regresión lineal multivariante generado para los alumnos de la titulación de ITIS se muestra en la Figura 2. Dicho modelo debe interpretarse de manera que determinados valores de ciertos atributos (que se indica a continuación del nombre del atributo para los atributos categóricos, y separado de éste por un guión) hacen que el rendimiento de un alumno varíe en un determinado valor, dado por el coeficiente de cada atributo del modelo, y cuyo signo indica si el rendimiento aumenta debido a esa característica, o disminuye. Además, también aparece un término independiente de cualquier atributo.

```
Rendimiento =
      D_Altr Estud_1 (21)· 35,55 +
      D_Estudis M_3 (140)· 4,063 +
      Ing Est_10 (248)· 9,673 +
      Ing Est_4 (3)· -36,54 +
      Ing Nota (572)· 7,572 +
      Ocupacio M_8 (73)· 6,573 +
      -27,52
```

Figura 2. Modelo de regresión para la titulación ITIS

En cada atributo se muestra, entre paréntesis, el número de alumnos afectados por dicho factor, en este caso, sobre el total de la muestra estudiada. El orden en el que aparecen los atributos en el modelo no presupone mayor o menor influencia en el rendimiento. El RECM asociado al modelo es 17,45.

Finalmente, para verificar la robustez de los resultados, los modelos se generaron dos veces, con diferentes particiones de los datos en conjuntos de entrenamiento y test realizadas aleatoriamente, y se comprobó que los modelos inducidos eran parecidos. También se calculó el rendimiento medio del total de alumnos estudiados, resultando un valor de 30,7, con una desviación típica de 19,6.

## 2.6. Interpretación de los resultados

Como puede observarse, los dos modelos mostrados en la sección anterior tienen un error similar. Sin embargo, no tienen en cuenta los mismos atributos, algo que no parece descartable por el hecho de haber utilizado técnicas diferentes. Incluso, dentro de un mismo modelo (por ejemplo, el árbol de decisión), no todos los atributos tienen la misma importancia, existiendo algunos que ni siquiera se tienen en cuenta. Para determinar la relevancia de cada atributo con respecto al rendimiento hemos analizado por separado cada uno de los modelos.

Análisis del Árbol C&R. Ya que un atributo puede utilizarse a diferentes niveles del árbol (véase *Ing Nota* en la Figura 1) y además repetidamente, hemos calculado para cada atributo el número de ejemplos para los cuales dicho atributo se utiliza (si el atributo se utiliza varias veces para el mismo ejemplo se cuenta tantas veces como se use). El número de ejemplos viene anotado en el Clementine entre paréntesis (tal y como ya hemos comentado) y hace referencia al conjunto de entrenamiento. Con esto tenemos que un atributo es relevante si se utiliza para un número importante de casos. Este número absoluto, que llamamos I (por importancia), lo dividimos entre el número total de ejemplos utilizados para el entrenamiento, obteniendo un valor denominado IR (importancia relativa). Finalmente, sumamos todas las IR de cada atributo y normalizamos, obteniendo un valor IR2 tal que todos los IR2 para todos los atributos sumen 1 y nos muestre un valor de importancia que permita comparar más fácilmente entre diferentes árboles. La Tabla 1 incluye esta información para el árbol de la Figura 1.

| Atributo | I | IR | IR2 |
|---|---|---|---|
| *D_Altr Estud* | 440 | 1 | 0,343 |
| *Ing Nota* | 584 | 1,327 | 0,455 |
| *Edad Ingreso* | 406 | 0,922 | 0,316 |
| *D_Estudis P* | 82 | 0,186 | 0,063 |
| *D_Estudis M* | 47 | 0,106 | 0,036 |
| *Ocupacio A* | 85 | 0,193 | 0,066 |
| *Ing Est* | 78 | 0,177 | 0,060 |
| TOTAL | 1282 | 2,913 | 1 |

Tabla 1. Análisis de los atributos en el árbol de decisión para ITIS

Hemos considerado que un atributo es relevante si IR2 > 0,05. Esto quiere decir que ese atributo se utiliza, en promedio, al menos en un 5% de las decisiones. Como hay varias decisiones por caso (todas las condiciones hasta un nodo hoja), un valor del 5% indica que el porcentaje de ejemplos afectados por este atributo será, por lo general, sensiblemente mayor que un 5%. De acuerdo a este criterio, los atributos relevantes son: *D_Altr Estud*, *Ing Nota* y *Edad Ingreso*, y en mucha menor medida, *D_Estudis P*, *Ocupacio A* e *Ing Est*. Ya sólo falta analizar si estos atributos contribuyen positivamente o negativamente, viendo cómo afecta al rendimiento medio de las dos hojas hijas según la partición del atributo. En el caso de que un atributo aparezca varias veces, se debe ver si la manera de afectar es consistente o algunas veces afecta positivamente o negativamente. Nótese que el hecho de decir que un atributo (o los valores del mismo) afecta positiva o negativamente es según las ramas del árbol en las que aparece. Podría ocurrir que un atributo afectara positivamente (o negativamente) en general (para casi todos los ejemplos), pero en un subconjunto de ellos representados por una rama del árbol sucediera al revés. El análisis ha dado los siguientes resultados:

- Los mejores rendimientos se obtienen para el valor 1 del atributo *D_Altr Estud,* es decir, los alumnos que ya poseen estudios universitarios, aunque esta condición sólo la cumple un porcentaje relativamente pequeño de alumnos.
- El atributo *Ing Nota* (nota de ingreso) afecta positivamente.
- El atributo *Edad Ingreso* afecta negativamente (cuanto mayor es, peor rendimiento).
- El atributo *D_Estudis P* afecta positivamente para los valores 1 y 3 (padre sin estudios o estudios superiores), y negativamente para el valor 2 (padre con estudios equivalentes a bachillerato).
- El atributo *Ocupacio A* afecta positivamente para los valores 2 y 3 (alumnos con una ocupación inferior a 15 horas o que no realiza trabajo renumerado), y negativamente para el valor 1 (alumnos con una ocupación mayor o igual a 15 horas semanales).
- El atributo *Ing Est* afecta positivamente para el valor 10 (alumnos que acceden desde

bachillerato LOGSE con PAU), y negativamente para el valor 2 (alumnos que acceden con prueba de acceso pero no provienen de bachillerato LOGSE).

Debemos tener en cuenta que los últimos tres resultados no son tan generales y afectan a un número relativamente pequeño de alumnos (según los valores de IR2 mostrados en la Tabla 1, la relevancia de estos atributos está en el límite considerado).

Análisis de la regresión lineal. En este modelo aparece en primer lugar el atributo *D_Altr Estud* con el valor 1 (*D_Altr Est_1*), indicativo de que poseen ya estudios universitarios. Este factor hace que el rendimiento medio de estos alumnos se incremente en 35,6 puntos, el coeficiente positivo más grande, aunque el número de alumnos que cumple esta condición (21) es reducido.

Los 248 alumnos con el atributo *Ing Est_10* (alumnos procedentes de LOGSE con PAU), con un coeficiente de 9,7 positivo tienen rendimiento medio superior a la media. En cambio, aquellos individuos que tienen el atributo *Ing Est_4* (titulados universitarios), aparecen con un coeficiente muy negativo, (-36,5), aunque sólo 3 individuos, con edades de ingreso muy superiores a la media (entre 35 y 55 años) aparecen en esta situación. Probablemente, cargas familiares y de trabajo podrían explicar su rendimiento muy por debajo de la media.

El valor 3 del atributo *D_Estudis M* (alumnos cuyas madres poseen estudios medios o superiores), afecta positivamente a un número importante de individuos (140) aunque con un coeficiente pequeño (4,1). Este atributo está relacionado con el valor 8 del atributo *Ocupacio M* (alumnos cuyas madres no tienen un trabajo remunerado), con un coeficiente de 6,6 positivo. Interpretamos estos dos atributos como que aquellos alumnos cuyas madres tienen estudios y además tienen disponibilidad para atender a sus hijos, influyen positivamente en el rendimiento de estos. Es obvio que esta positiva influencia no tiene porqué haberse producido durante el primer año de estudios universitarios, sino durante toda la enseñanza previa.

Finalmente aparece el atributo numérico *Ing Nota* con un coeficiente de 7,6. En este caso, al ser este atributo numérico, el valor del coeficiente debe ser multiplicado por el valor del atributo, que por ser la nota de acceso debe ser siempre igual o superior a 5.

## 3. Conclusiones

Como primera conclusión, podemos afirmar que las técnicas de minería de datos proporcionan una herramienta que permite determinar qué características de los alumnos de nuevo ingreso son más relevantes de cara a estimar su rendimiento académico el primer año. La desviación típica obtenida con el simple cálculo de la media aritmética del rendimiento es algo mayor que en los modelos (19,6 vs 17,95 y 17,45), pero la posibilidad de establecer factores determinantes del rendimiento es, a nuestro entender, la mayor ventaja de esta técnica. En el caso estudiado, factores como los estudios previos del alumno y la nota de ingreso en la titulación aparecen de manera repetida como claramente correlacionados con el rendimiento académico el primer año. También aparecen factores que podrían influir en el rendimiento, como las ocupaciones y estudios de los padres, o la edad de ingreso del alumno, aunque estos dependen de la técnica utilizada. En cambio, el país de procedencia o el lugar de residencia (del alumno o de su familia) no aparece en ningún caso.

Si bien pretendemos que los resultados específicos de este estudio sean útiles de inmediato, este trabajo se enmarca dentro de un proyecto más general que ha realizado estudios en los últimos años acerca de diferentes indicadores (abandono, duración de estudios, ...) sobre diferentes titulaciones (II, ITIG, ITIS) y diferentes cohortes. Esta es la primera vez, sin embargo, en la que toda la información se integra adecuadamente en un almacén de datos, se aplican las técnicas de minería de datos de una manera sistemática y se realiza un análisis de los modelos extraídos por equipos multidisciplinares. Estos equipos conjugan los conocimientos técnicos sobre análisis de datos con una gran experiencia en la docencia universitaria, incluyendo la participación y el apoyo de las direcciones de los centros en este análisis, así como en el despliegue y aplicación del conocimiento y de los modelos extraídos. Pensamos que este compromiso y esta continuidad es fundamental de cara a amortizar el esfuerzo realizado en las primeras etapas del

proceso: establecimiento de objetivos, entendimiento de los datos, así como de la siempre compleja limpieza e integración de datos.

En el futuro inmediato planteamos extender el análisis a otros cursos, posiblemente utilizando otros indicadores (abandono, duración de estudios), desglosar el análisis por asignaturas o por temáticas, así como ampliar las herramientas empleadas (por ejemplo *SPSS o R-project, Weka*). También estamos actualmente trabajando en un proyecto para contrastar resultados de rendimiento académico en estudios de informática en colaboración con otras universidades españolas.

No obstante, creemos que un programa estable de análisis de datos que cuente con el apoyo institucional de los centros involucrados y de la propia universidad sería fundamental para poder diseñar estrategias de mejora más adecuadas a la realidad de cada centro.

## Agradecimientos

A la Universidad Politécnica de Valencia por la financiación recibida a través del programa PACE. A las direcciones de la FI y la ETSIAp, por facilitarnos los datos.

A Lorenzo Morales, subdirector del Área de Sistemas de Información y Comunicaciones, que amablemente nos resuelve todas nuestras dudas acerca de la base de datos de la universidad, y a Sara Collado, jefa de administración de la ETSIAp, que nos da luz sobre todas nuestras dudas administrativas.

## Referencias

[1] Bará, J., y otros. Informe Transversal del rendimiento académico de las ingenierías técnicas. Consejo de Universidades, 2001.

[2] Breiman, L., Friedman, J., Losen, R., Stone, *Classification and Regression Trees*. Wadsworth and Brooks, 1984 (new edition 1993).

[3] Fontes, R. y otros. *Valor predictivo de algunos criterios de selección para el ingreso a la carrera de medicina*, Revista Cubana de Educación Médica Superior, vol.14, no.1, p.17-25. 2000.

[4] Hernández, J., Ramírez, M.J., Ferri, C. *Introducción a la Minería de Datos* Prentice Hall, 2004.

[5] http://www.mec.es ⇒ Universidades ⇒ Consejo de Coordinación Universitaria⇒ Directrices para la elaboración de títulos universitarios (Enero 2007).

[6] http://www.mec.es ⇒ Universidades ⇒ Estadísticas ⇒ Datos y Cifras del Sistema Universitario Español (Enero 2007).

[7] http://www.mec.es ⇒ Universidades ⇒ Estadísticas ⇒ Indicadores ⇒ Catálogo de Indicadores (Enero 2007).

[8] http://www.spss.com/clementine (Enero2007)

[9] Muñoz, S. *Indicadores de rendimiento académico del alumnado de la Universidad de La Laguna*, Jornadas sobre Políticas de Calidad en la Universidad de La Laguna, 2005.

[10] Zúnica, L., Blesa, P., Alcover, R., Más, J., Valiente., J. *Estudio del rendimiento académico de una asignatura con relación a asignaturas de cursos anteriores*, Libro de resúmenes JENUI 2003, p. 137-142, Cádiz, 2003.