Actas de las Jenui, vol. 1. 2016.
Páginas: 47-54

Redes bayesianas para identificar perfiles de estudiante. Aplicación al estudio del abandono de las titulaciones de Informática en la Universidad de Castilla-La Mancha

Carmen Lacave, Ana I. Molina, Miguel A. Redondo, Manuel Ortega
Dpto. de Tecnologías y Sistemas de la Información, UCLM
{carmen.lacave, anaisabel.molina, miguel.redondo, manuel.ortega}@uclm.es

Resumen

El abandono de los estudiantes es un problema que afecta a todas las universidades siendo más acusado en las titulaciones de las ramas de Ingeniería y Arquitectura. Como docentes del Grado de Ingeniería Informática de la Universidad de Castilla-La Mancha nuestro interés se centra en analizar el perfil del estudiante que abandona estos estudios, con el fin de definir acciones orientadas a reducir la actual tasa de abandono. En ediciones anteriores de las JENUI se ha analizado esta problemática desde el punto de vista de la estadística tradicional y de la minería de datos, mediante árboles de decisión y regresión multivariante; en este trabajo lo abordamos mediante algoritmos de aprendizaje de redes bayesianas, ya que éstas tienen una semántica muy rica y son fácilmente interpretables. Los resultados del trabajo no son concluyentes debido a las restricciones de la base de datos utilizada, pero la descripción del estudio realizado pone en valor el interés de la técnica empleada y sienta las bases para mejorar el alcance de la investigación en trabajos futuros relacionados con la extracción de perfiles de estudiantes.

Abstract

Student dropout is a problem that affects all universities although it is more significant in Engineering and Architecture. Since we are teachers of the Degree of Informatics at University of Castilla-La Mancha, our interest is to analyse the profile of the student who abandons these studies in order to define actions to reduce the dropout rate. In previous editions of JENUI this problem has been analysed from the point of view of traditional statistics and data mining techniques based on decision trees and multivariate regression; in this work we use learning algorithms for Bayesian networks because these have a rich semantic and are easily interpretable. The conclusions of the study are limited because of the database used but the article reflects the interest of the technique applied and lays the foundations for improving the scope of the investigation in future works related to the extraction of profiles of different types of students.

Palabras clave

Minería de datos, perfiles de estudiante, abandono, estudios Informática, redes bayesianas.

1. Introducción

Una de las responsabilidades más importantes que tiene cualquier estado con sus ciudadanos es ofrecer una educación de calidad, lo que implica no sólo altos niveles de producción de conocimiento, sino que la educación sea impartida de manera eficiente para que los estudiantes puedan aprender sin ningún problema [15]. Una forma de conseguirlo es que las instituciones educativas conozcan las características y dificultades de sus alumnos de manera que puedan tomar medidas correctoras, con altas probabilidades de éxito si se aplican a los estudiantes de acuerdo a su perfil. En este contexto, el abandono de los estudios universitarios es una realidad que afecta a todas las universidades [17] y conlleva pérdidas económicas, problemas sociales y posibles problemas psicológicos en el estudiante [11]. Este problema afecta de modo significativo a las titulaciones relacionadas con la Ingeniería Informática [5, 10, 13, 15] lo que ha motivado su estudio desde distintas perspectivas y haciendo uso de técnicas diferentes. Así, hay trabajos que se basan en la creación de encuestas o cuestionarios a responder por los alumnos y cuyos datos son posteriormente analizados mediante estadísticos descriptivos [5]. En otros, estas técnicas se han aplicado sobre los datos suministrados por las propias universidades [10]. Sin embargo, existe otro enfoque muy interesante, y más actual, basado en el uso de técnicas de *minería de datos* [15]. De este modo, suponiendo que los datos contienen más información oculta de la que se ve a simple vista, la minería de datos surge como un conjunto de técnicas que pueden ser usadas para extraer conocimiento relevante e interesante de los

mismos [7, 19]. Haciendo uso de árboles de decisión y/o regresión multivariante, se han obtenido indicadores de rendimiento o factores de riesgo de abandono en distintas universidades, españolas [1] y del resto del mundo [2, 11, 15]. Uno de los principales inconvenientes de estos métodos es la dificultad para interpretar los resultados a través de los modelos obtenidos. Sin embargo, las redes bayesianas [16] aportan ventajas frente a las demás gracias a su rica semántica, que permite al usuario entender fácilmente los resultados, por lo que se han utilizado con éxito para la extracción de perfiles educativos [9, 14]. Una red bayesiana se define por un grafo dirigido acíclico, en el que cada nodo representa una variable aleatoria, y los enlaces representan las dependencias probabilísticas entre las variables. Además, cada nodo lleva asociada la distribución de probabilidad de dicho nodo condicionada en sus padres, que definirá, junto con la estructura del grafo, la dependencia entre los mismos.

En el caso concreto de la Universidad de Castilla-La Mancha (UCLM) no tenemos constancia de que se hayan realizado estudios sistematizados y rigurosos que evalúen el perfil del estudiante que abandona[1] alguna de las titulaciones de Informática. Sin embargo, los porcentajes de abandono son preocupantes, situándose en el caso de nuestra universidad en torno al 40%. Por ello nos hemos planteado identificar el perfil del alumno que deja los estudios de Ingeniería Informática aplicando técnicas de minería de datos basadas en el aprendizaje automático de redes bayesianas.

Para seguir sin problemas la lectura del artículo, en la próxima sección se presentan los fundamentos teóricos en los que se basa la metodología utilizada, que es descrita con detalle en la sección siguiente. A continuación se discuten los resultados más importantes y, por último, se presentan las principales conclusiones así como las limitaciones más significativas.

2. Fundamentos teóricos

En esta sección explicamos brevemente los conceptos teóricos para entender el proceso de aprendizaje de una red bayesiana a partir de una base de datos.

2.1. Introducción a la minería de datos

El proceso de extracción de conocimiento a partir de bases de datos, conocido como KDD (*Knowledge Discovery in Data Bases*), es "un proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y entendibles sobre un conjunto de datos, normalmente mediante el uso de técnicas de inteligencia artificial, cuyo objetivo es el de encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad" [7]. En este contexto, los *datos* se refieren a un conjunto de hechos (ejemplos en una base de datos) y los *patrones* son resultados o expresiones en algún lenguaje que describen de manera compacta dichos datos. Finalmente, el término *no trivial* significa que se lleva a cabo algún proceso de búsqueda o inferencia, es decir, involucra la búsqueda de estructuras, modelos, patrones o parámetros. El proceso KDD consta de varias etapas: selección, pre-procesamiento y transformación de los datos iniciales, la aplicación posterior de alguna técnica normalmente basada en inteligencia artificial a los datos procesados y, finalmente, la interpretación y/o evaluación de los patrones obtenidos en la etapa anterior, con la consiguiente obtención de conocimiento. La etapa de aplicación de algoritmos que construyen modelos a partir de los datos, con el fin de describir o predecir el comportamiento de un cierto fenómeno es lo que se conoce como *minería de datos*. Sin embargo, generalmente se asocia el concepto de minería de datos a todo el proceso de KDD, en lugar de a la etapa de extracción de conocimiento.

Una de las diferencias que existe entre el análisis de datos tradicional (normalmente basado en el uso de técnicas estadísticas) y la minería de datos, es que el primero supone que las hipótesis ya están construidas y validadas sobre los datos, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos; es decir, que se descubre la información sin necesidad de formular previamente una hipótesis. Ello se logra a través de la aplicación automatizada de algoritmos, normalmente de inteligencia artificial, que permiten detectar modelos o patrones en los datos, lo cual es más interesante que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad.

En el contexto educativo, la minería de datos se puede utilizar para la búsqueda, análisis y la extracción de patrones de conocimiento, que permitan mejorar el proceso enseñanza-aprendizaje a partir de modelos predictivos, de forma cualitativa y cuantitativa [3].

[1] Los datos que nos proporcionó la OPyC reflejaban que el "estudiante que abandona" es aquel alumno de alguna de las titulaciones de Informática de la UCLM que no vuelve a matricularse en su titulación ni en ninguna de las titulaciones de Informática de la UCLM en el año siguiente. Somos conscientes de que esta definición no concuerda con muchas de las definiciones que existen y que consideran al alumno que abandona aquel que no se matricula en los dos años siguientes al último. A pesar de ello, y ante la imposibilidad de obtener una nueva base de datos, decidimos continuar con nuestro trabajo.

2.2. Redes bayesianas

Una *red bayesiana* es un modelo gráfico probabilístico que permite representar relaciones de dependencia e independencia probabilística entre una colección de datos. Se define mediante un grafo dirigido acíclico, en el que cada nodo representa una variable y cada arco una dependencia probabilística, que viene dada por la probabilidad condicional de cada variable, dados sus padres [16]. Las variables pueden ser discretas o continuas, pero solo vamos a considerar aquí las del primer tipo. La Figura 1 muestra un ejemplo muy sencillo de una red bayesiana, con cinco nodos binarios (valores de tipo Sí/No), que representan el *Cáncer de Pulmón*, las causas que lo provocan (ambientes con *Polución* y ser *Fumador*), y un síntoma (*Disnea*) y un signo (los *Resultados* de la prueba de *Rayos X*) que permiten su diagnóstico. Los arcos representan las relaciones causa-efecto entre dichos nodos, definidas por la probabilidad de todos los valores de cada nodo, condicionada en las distintas configuraciones de valores de sus padres. Por ejemplo, el valor P(C=s|P=b, F=n)=0,001 indica que la probabilidad de que un paciente padezca cáncer de pulmón (C=s), sabiendo que no hay polución (P=b) y que no es fumador (F=n), es de 0,001; lo que se puede interpretar también como que 1 de cada 1000 personas que no son fumadoras ni viven en ambientes contaminados padecen cáncer de pulmón.

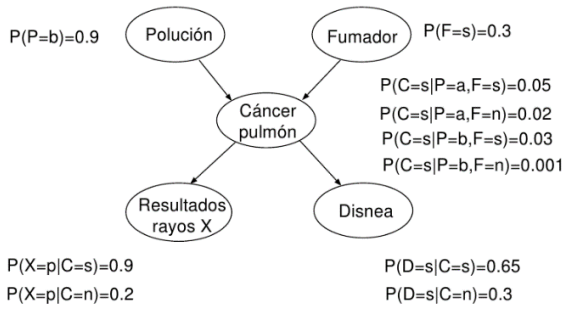

Figura 1. Ejemplo de red bayesiana que representa las causas, síntomas y signos del cáncer de pulmón.

Así pues, las redes bayesianas proporcionan una forma compacta de representar el conocimiento. Además permiten combinar el conocimiento dado por un experto humano con los datos recogidos en una base de datos, incluso teniendo éstas datos incompletos, y facilita la construcción de modelos que se ajustan a la realidad evitando el sobreajuste que se puede producir por el empleo de datos que representen sólo una parte de la realidad. En consecuencia, son una muy buena alternativa para su uso en la minería de datos porque muestran las relaciones de dependencia entre las variables. También proveen métodos flexibles de razonamiento basados en el Teorema de Bayes[2] (de ahí su nombre), y bien fundamentados en la teoría de la probabilidad, capaces de predecir el valor de variables no observadas y explicar las observadas. Se puede razonar sobre una red de dos formas [16]:

- Obteniendo las probabilidades *a posteriori*[3] de las variables de interés, dado que se conoce el valor que toman algunas otras variables observadas. Este tipo de razonamiento se suele utilizar en sistemas donde se desee realizar un diagnóstico o una predicción.
- Buscando la configuración de las variables que maximicen la probabilidad conjunta dada la evidencia observada. Este proceso se conoce como *abducción* y se utiliza para explicar la evidencia observada.

Una red bayesiana se puede construir de forma manual, mediante la ayuda de expertos en el dominio a modelar, o de forma automática, mediante la aplicación de algoritmos de aprendizaje [16], que puede ser de dos tipos:

- *Aprendizaje estructural*: se obtiene la estructura de la red bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas.
- *Aprendizaje paramétrico*: dada una estructura, proporciona las probabilidades a priori y condicionales a partir de una estructura dada.

Existen varios algoritmos para construir una red bayesiana a partir de una base de datos, siendo los más conocidos el K2 [6] y el PC [18]. El algoritmo K2 intenta encontrar una red óptima en términos de la verosimilitud de la base de datos para cada red candidata. En cambio el algoritmo PC trata de determinar la estructura de la red a través de pruebas estadísticas de independencia. Ninguno de los dos métodos es absolutamente superior al otro y ambos operan con variables discretas.

3. Metodología

Las etapas seguidas durante nuestro trabajo, descritas en las secciones siguientes, fueron:

1. Definición del objetivo del trabajo.
2. Establecimiento de la población de estudio.
3. Preparación de la base de datos para eliminar anomalías que proporcionen resultados de calidad.

[2] El Teorema de Bayes permite obtener la probabilidad de un suceso aleatorio *A* dado *B*, *P(A|B)*, en términos de la distribución de probabilidad condicional del suceso *B* dado *A*, *P(B|A)*, y la probabilidad de A, *P(A):P(A|B)=P(B|A)*P(A)/P(B)*.

[3] En estadística bayesiana, la *probabilidad a posteriori* de un evento aleatorio A es la probabilidad condicional que se obtiene al conocer una evidencia E, es decir, *P(A|E)*, y se suele denotar por *P*(A)*.

4. Elección y aplicación de las técnicas de minería de datos adecuadas.
5. Interpretación del modelo para la discusión posterior de los resultados.

3.1. Objetivo del trabajo

Como hemos indicado previamente, nuestro objetivo inicial era el de obtener el perfil del alumno que abandona alguna de las titulaciones de Informática. Sin embargo, entrevistar a quienes ya habían abandonado suponía una tarea que, además de larga, difícil y laboriosa, podía entrar en conflicto con cuestiones legales de protección de datos. Teniendo en cuenta que las redes bayesianas se han utilizado con éxito para obtener el perfil de un estudiante a partir de bases de datos de la universidad [9, 14] nos planteamos la siguiente pregunta de investigación:

P1: *¿Cuál es el perfil que proporciona la base de datos de la UCLM del alumno que abandona alguna de las titulaciones de Ingeniería Informática de dicha universidad?*

La búsqueda de respuesta a esta pregunta constituye el resto del trabajo descrito a continuación.

3.2. Definición de la población de estudio

Para ello, la Oficina de Planificación y Calidad de la UCLM (OPyC) nos proporcionó parte de los datos que les habíamos solicitado de los alumnos de las titulaciones de Ingeniería Informática (ISI), Ingeniería Técnica en Informática de Gestión (ITIG), Ingeniería Técnica en Informática de Sistemas (ITIS) y el Grado en Ingeniería Informática (GII), impartidas tanto en el Campus de Ciudad Real como en el Campus de Albacete de la UCLM y que habían abandonado en alguno de los cursos del 2008-2009 al 2011-2012.

3.3. Preparación de la base de datos

La base de datos facilitada contenía 491 registros con 18 campos correspondientes a los datos de los alumnos que abandonaron en algún año académico dentro del período 2008-2009 a 2011-2012. Se eliminaron los que cambiaron a otra carrera de Informática y a los que vienen a la UCLM a estudiar temporalmente (Erasmus y Sicue/Séneca). El listado resultante contenía alumnos duplicados, por ejemplo aquellos que se matriculan un año, dejan de estudiar durante los dos siguientes y vuelven a matricularse en el último. Además, había registros correspondientes a varias matrículas de un alumno en el mismo año. También aparecían datos erróneos, como un alumno con nota de acceso igual a cero. Esta multiplicidad de datos, junto con algunos datos ausentes, como por ejemplo los estudios o la profesión de los padres en algunos casos, se pudieron corregir haciendo uso de otra base de datos que nos proporcionó la OPyC de 6707 registros, con 25 atributos cada uno, extraída de las matrículas en el período 2008-2009 a 2012-2013. Al no contar con ayuda supervisada por algún miembro de la OPyC, el "cruce" de ambas bases de datos permitió eliminar todas estas anomalías y supuso la reducción de la base de datos inicial a 363 registros.

Además, definimos las siguientes 18 variables, una por campo, que representan:

- *DESC_PLAN,* el nombre de la titulación de Informática que se abandona.
- *DESC_CENTRO*, el nombre del centro en el que se ha matriculado el alumno.
- *SEXO*, el sexo del alumno que abandona.
- *EDAD* del alumno cuando abandonó sus estudios (coincide con la edad de su última matrícula).
- *DESC_PROVINCIA_FAM*, la provincia de residencia familiar.
- *DESC_MUNICIPIO_FAM*, el municipio de residencia familiar.
- *TIPO_ACCESO,* la forma de acceso a la titulación.
- *SUBTIPO_ACCESO,* diversos tipos de valores en función del tipo de acceso.
- *NOTA_ACCESO*, la nota con la que se accedió a los estudios, en su caso.
- *ESTUDIOS_PADRE*, el nivel de estudios del padre.
- *ESTUDIOS_MADRE:* ídem para la madre.
- *PROFESION_PADRE*, la cualificación de la profesión del padre.
- *PROFESION_MADRE*: ídem para la madre.
- *#ASIG_MATRICULADAS*: Número total de asignaturas matriculadas del curso 2008-09 al 2012-13.
- *#ASIG_APROBADAS*: Ídem para las asignaturas aprobadas.
- *#ASIG_SUSPENSAS*: Ídem para las asignaturas suspensas.
- *#ASIG_CONVALIDADAS:* ídem convalidadas.
- *#ANNOS(ABANDONO-PRIMERA_MATRICULA)*, diferencia en años entre el curso de abandono y el de la primera matrícula. La base de datos incluía alumnos que se matricularon por primera vez hasta 13 años antes del año en el que abandonaron la titulación, de ahí que algunos resultados posteriores puedan llamar la atención.

Esta etapa de preparación de la base de datos ha sido la más laboriosa, lo que confirma lo que ya ha sido puesto de manifiesto por otros autores [1]. El proceso

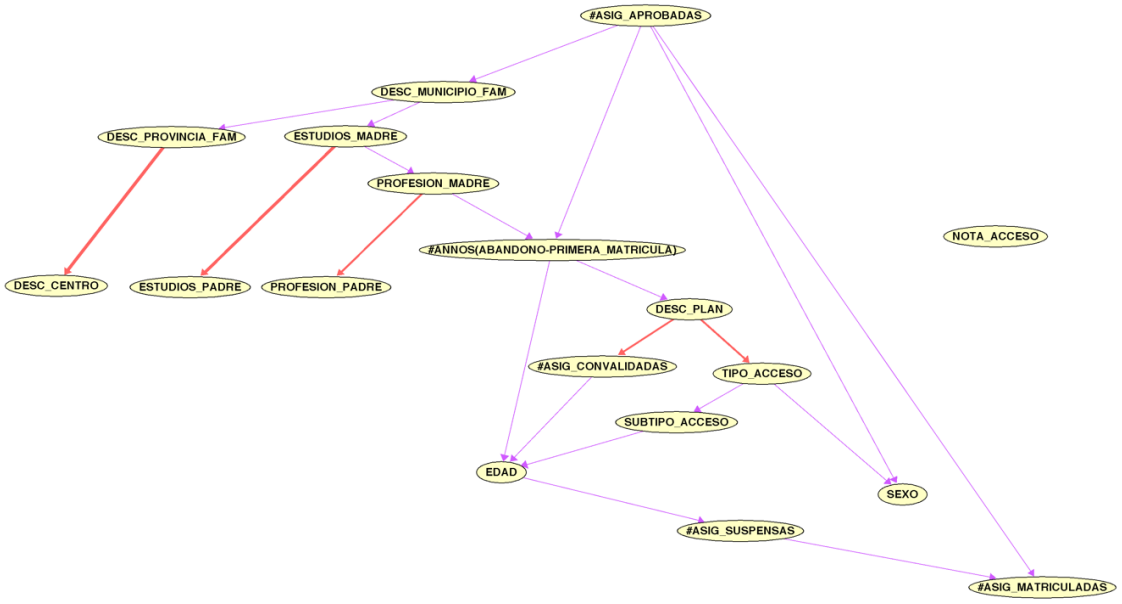

Figura 2. Red Bayesiana obtenida mediante la aplicación del algoritmo de aprendizaje K2 a la base de datos

se llevó a cabo con el sistema administrador de base de datos Microsoft SQL Server, versión Developer.

3.4. Minería de Datos

De entre las técnicas existentes de minería de datos, hemos utilizado la de aprendizaje automático de una red bayesiana, por la potencia semántica que proporcionan este tipo de modelos, entre otras cosas, tal y como se indicó en las secciones anteriores. Además, los algoritmos de abducción se pueden aplicar perfectamente a la búsqueda del perfil más probable de los individuos de una población bajo determinadas condiciones impuestas por las variables observadas [14]. Sin embargo, uno de los inconvenientes que presenta este tipo de modelos gráficos es que el exceso de granularidad en la definición de los valores de las variables puede aumentar la complejidad, no solo del modelo obtenido, sino del coste de ejecución de los propios algoritmos, tanto de los de aprendizaje como de los de razonamiento posterior. Por tanto, como paso previo, se procedió a simplificar algunas variables que tenían demasiados valores, como las relacionadas con la edad, el municipio del domicilio familiar, los estudios y las profesiones paternas, y las relacionadas con el total de asignaturas (matriculadas, convalidadas, etc.). Por ejemplo, para la descripción del municipio familiar se decidió analizar únicamente si el alumno provenía de un pueblo o de una capital de provincia; los valores de la edad se agruparon por intervalos, así como el número de asignaturas o el número de años que el alumno permanece en la titulación antes de abandonarla, etc. En definitiva, los valores de cada variable fueron los siguientes:

- *DESC_PLAN:* ISI, ITIG, ITIS y GII.
- *DESC_CENTRO*: Ciudad Real o Albacete.
- *SEXO*: Hombre o Mujer.
- *EDAD:* [20,25], [26,30], [31,40], [41-50].
- *DESC_MUNICIPIO_FAM:* Pueblo, Capital de Provincia.
- *TIPO_ACCESO* a los estudios: Acceso a ciclos, Convalidación de estudios extranjeros, Preinscripción en 1º curso, Traslados en 2º y posteriores o Traslados con preinscripción en 1º.
- *SUBTIPO_ACCESO*: COU, Diplomado, FPII, etc.
- *NOTA_ACCESO*: [5,6), [6,7), [7,8), [8,9), [9,10].
- *ESTUDIOS_PADRE, ESTUDIOS_MADRE*: Sin Estudios, Primarios, Secundarios, Superiores.
- *PROFESION_PADRE*, *PROFESION_MADRE*: valores del 0 al 10 que representan distintos niveles profesionales, desde parado (0) a director de una gran empresa (10).
- *#ASIG_MATRICULADAS*, *#ASIG_APROBADAS*, *#ASIG_SUSPENSAS*, *#ASIG_CONVALIDADAS:* [1-5], [6-10], [11-15], [16-20], [21-25], [26-30], [31-35].
- *#ANNOS(ABANDONO-PRIMERA_MATRICULA)*: 1, [2-3], [4-6], [7-13].

A partir de la base de datos resultante, se realizó el aprendizaje de la red mediante la aplicación del

algoritmo K2 y la herramienta OpenMarkov[4] [4]. El uso de este programa se explica por varias razones: es gratuita y de código abierto, incluye la opción de ejecutar diversos algoritmos de aprendizaje paso a paso, permite la edición de la red bayesiana en cualquier momento, ofrece opciones de pre-procesamiento de datos y consta de una interfaz que facilita todo el proceso. La red obtenida se muestra en la Figura 2 y en ella se pueden ver las dependencias e independencias existentes entre las 18 variables involucradas (véase sección 3.3.).

3.5. Interpretación del modelo

Aunque la herramienta OpenMarkov es muy útil para el aprendizaje de redes bayesianas, para la etapa de interpretación del modelo se utilizó la herramienta Elvira[5] [8] porque proporciona opciones de explicación, tanto del modelo como del razonamiento [12], que facilitan la comprensión de los resultados obtenidos. Ya que los enlaces de la red obtenida no se deben interpretar como relaciones causales, sino como dependencias probabilísticas, resulta muy útil contar con una herramienta que ayude a interpretar el modelo. En el caso de Elvira, el tipo de dependencia entre las variables se ilustra mediante el coloreado y el grosor de los enlaces [12]. Para ello, es necesario que los valores de los nodos que definen la red bayesiana estén ordenados y que ésta esté almacenada en formato compatible con Elvira. Ambos requisitos se pueden garantizar haciendo uso de OpenMarkov. La Figura 2 muestra la red abierta en la herramienta Elvira, en la que se reflejan las dependencias e independencias probabilísticas entre las variables. Por ejemplo, el centro en el que se matricula el alumno depende únicamente de la provincia de residencia familiar; el número de años que pasa el alumno en la titulación hasta que abandona depende del número de asignaturas aprobadas y de la profesión de la madre[6]. El grosor de cada enlace es proporcional a la influencia que cada nodo transmite a otro y el color rojo (o más oscuros) representan dependencias probabilísticas positivas [12], es decir, que cuanto mayores sean los valores que toma el padre aumenta la probabilidad de que el hijo tome valores mayores; los enlaces en violeta (más claros) indican que la influencia es desconocida, es decir, que para ciertos valores del padre hay influencia positiva y para otros, negativa.

Así, el enlace entre las variables *ESTUDIOS_MADRE* y *ESTUDIOS_PADRE* refleja que existe una relación positiva entre los estudios de la madre y los del padre. De forma similar, el enlace rojo entre *PROFESION_MADRE* y *PROFESION_PADRE* indica que a mayor nivel de la profesión de la madre mayor será el nivel de la profesión del padre. Por otra parte, la titulación que se abandona (*DESC_PLAN*) correlaciona positivamente con el tipo de acceso a los estudios (*TIPO_ACCESO*) y el número de asignaturas convalidadas. Además, la provincia de procedencia del alumno que abandona (*DESC_PROVINCIA_FAM*) influye directamente en la elección del centro de estudios (*DESC_CENTRO*), lo que era de esperar.

Un resultado curioso es que la nota de acceso a la titulación (*NOTA_ACCESO*) es independiente de todas las características que hemos tenido en cuenta en este trabajo, ya que no aparece ningún enlace que relacione dicha variable con alguna de las demás. Esto tiene cierto sentido pues la nota media es de 5,6 (desviación típica=0,85). No obstante, es posible que al aumentar el tamaño de la base de datos desaparezca esta independencia.

4. Resultados

A partir de la red aprendida, y haciendo uso de la herramienta Elvira, se procedió a aplicar un proceso de *abducción total* para obtener la configuración más probable para todas las variables que forman parte de la red que responde a la pregunta de investigación **P1** (sección 3.1). Así, se obtuvo que el perfil más probable del alumno que abandona alguna de las titulaciones de Informática en la UCLM, con una probabilidad de 0,000082, corresponde al de varón, de 31 a 40 años de edad, matriculado en Ingeniería Técnica en Informática de Sistemas de Ciudad Real, procede de un pueblo de la provincia de Ciudad Real, accedió a la universidad por preinscripción en primer curso y selectividad LOGSE/LOE; la nota con la que accedió fue con un 5, los estudios de los padres son de nivel primario, la profesión del padre es del grupo "servicios de restauración, personales, protección y vendedores de comercio"; la profesión de la madre es del grupo de "operador de instalaciones y maquinaria"; el total de asignaturas aprobadas fue menor que 5, y estuvo en la titulación que abandonó de 7 a 13 años. El segundo perfil más probable, con probabilidad de 0,000074, únicamente se diferencia de éste en que el centro de estudios corresponde a Albacete y procede de dicha ciudad.

Estas probabilidades tan bajas revelan una gran heterogeneidad de los datos estudiados, debido a la elevada granularidad de algunas variables, como la provincia de residencia familiar, las profesiones paternas, el nivel de estudios, etc. Por otra parte, las probabilidades obtenidas reflejan que el modelo no es demasiado preciso, ya que el perfil más probable no corresponde a ningún registro de la base de datos. Esto indica que el número de registros de la base de datos no permite ajustar con precisión el modelo.

[4] Disponible en www.openmarkov.org
[5] Disponible en leo.ugr.es/elvira

[6] Debido a la limitación de espacio no nos es posible ahondar más en la semántica del grafo.

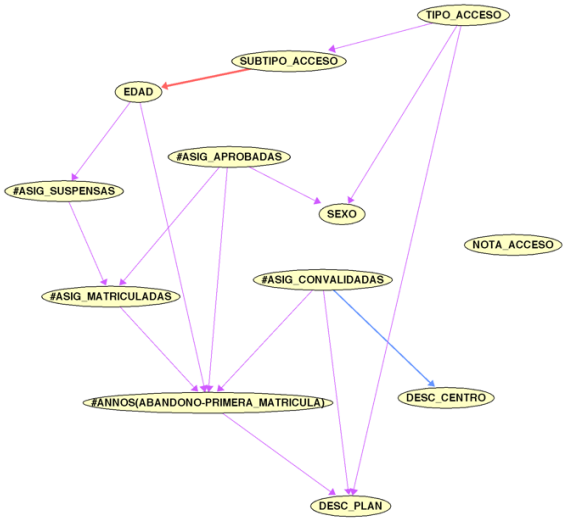

Figura 3. Red Bayesiana obtenida mediante la aplicación del algoritmo de aprendizaje K2 a la base de datos simplificada.

En consecuencia, decidimos simplificar nuestro objetivo, tratando de identificar el perfil del alumno en base únicamente a sus datos académicos. Por tanto, seleccionamos de la base de datos las 12 variables siguientes: *DESC_PLAN, DESC_CENTRO*, *SEXO*, *EDAD, TIPO_ACCESO*, *SUBTIPO_ACCESO*, *NOTA_ ACCESO*, *#ASIG_MATRICULADAS*, *#ASIG_APROBADAS*, *#ASIG_SUSPENSAS*, *#ASIG_CONVALIDADAS* y *#ANNOS(ABANDONO-PRIMERA_MATRICULA)*.

En la Figura 3 se muestra la red aprendida a partir de los 363 registros de la base de datos y los 12 campos correspondientes a las 12 variables seleccionadas. Se observa ahora una relación negativa entre *#ASIG_CONVALIDADAS y DESC_ CENTRO*. Como en el orden (aleatorio) de los valores de esta variable el centro de Albacete aparece antes que el de Ciudad Real, el enlace indica que a mayor número de asignaturas convalidadas aumenta la probabilidad de que la titulación abandonada pertenezca al campus de Albacete. Igual que en el modelo anterior, en éste la nota de acceso (*NOTA_ACCESO*) vuelve a ser independiente del resto de las variables. Este hecho nos plantea el interrogante, de cara a trabajos futuros, de si esta variable influye en el abandono del alumno.

Sobre este modelo nos planteamos la siguiente preguntas de investigación:

P2: *¿Cuál es el perfil del alumno que abandona y qué titulación, si solo se tienen en cuenta el sexo, la edad y sus datos académicos?*

Para responder a esta pregunta se aplicó nuevamente un proceso de *abducción total* sobre todas las variables que forman parte de la red de la Figura 3. Así, se obtuvo que el perfil más probable del alumno que abandona alguna de las titulaciones de Informática en la UCLM, con una probabilidad de 0,01 (1% de los alumnos), corresponde al de varón, de 20 a 25 años de edad, matriculado en el Grado de Informática de Ciudad Real, accedió a la universidad por preinscripción en primer curso y selectividad LOGSE/LOE; la nota con la que accedió fue con un 5, el total de asignaturas aprobadas fue menor que 5, el número de asignaturas convalidadas fue también menor que 5 y estuvo en la titulación que abandonó 1 año.

Estos resultados están más cercanos de la realidad, aunque siguen sin reflejarla con exactitud, ya que en la base de datos hay 8 alumnos con estas características, lo que equivale a una frecuencia de aparición del 2,3%.

5. Conclusiones

En este trabajo se ha tratado de identificar el perfil del estudiante que abandona la titulación de Informática en la UCLM, a partir de datos de las matrículas, mediante el uso de redes bayesianas. Este enfoque constituye una propuesta novedosa respecto a trabajos previos que han abordado el mismo problema haciendo uso de técnicas de minería de datos. La principal ventaja del uso de redes bayesianas reside en que el modelo gráfico es fácil de interpretar. Pero además, los algoritmos de abducción permiten la obtención de perfiles (parciales y totales).

A lo largo de este proceso, la etapa de preparación de la base de datos fue la más larga y laboriosa. La obtención de la red ha sido posible gracias al uso de OpenMarkov, que facilita el aprendizaje y la edición posterior de la red obtenida, y de Elvira, que ayuda a interpretar los resultados.

Entre las principales limitaciones del estudio hay que destacar que no se pudo contar con los datos de abandono de los años 2012-2013 al 2014-2015 ni los previos a 2008-2009, lo que hubiera permitido obtener un modelo más preciso a partir del cual determinar el perfil de alumno que ha abandonado hasta ahora. Por otra parte, los años que se han estudiado han coincidido con épocas de cambio de planes de estudio y con el cambio a los grados, lo que ha aumentado la heterogeneidad de los datos.

En consecuencia, los resultados obtenidos no permiten tomar decisiones en relación al problema planteado (preguntas P1 y P2). Sin embargo, la metodología descrita se puede tomar como referencia para otros trabajos en los que se cuente con una base de datos más completa y se desee conocer el perfil de los alumnos que merezcan ser tenidos en cuenta de cara a la toma de decisiones por parte de la universidad.

Finalmente, como trabajo futuro nos planteamos utilizar las redes bayesianas para identificar las variables más relevantes que permitan predecir el riesgo de abandono, ya que no ha sido posible abordarlo hasta ahora y tanto los datos como el modelo generado deben ser distintos.

Este trabajo ha sido financiado parcialmente por el proyecto del MINECO TIN2015-66731-C2-2-R, por la Red Iberoamericana CYTED (Red 513RT0481), y los proyectos de la Junta de Comunidades de Castilla-La Mancha PPEII-2014-012A y PPII11-0013-1219. Agradecemos a la Oficina de Planificación y Calidad de la UCLM su colaboración desinteresada.

Referencias

[1] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica, "Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos", Actas de las JENUI 07, pp. 163-170, 2007.

[2] Kamagate Azoumana, "Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos", *Revista Pensamiento Americano*, 6 (10), pp. 41-51, 2013.

[3] Alejandro Ballesteros, Daniel Sánchez-Guzmán, Ricardo García, "Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo", *Latin-American Journal of Physics Education,* 7(4), pp. 662-668, 2014.

[4] Íñigo Bermejo, Jesús Oliva, Franciso J. Díez y Manuel Arias, "Interactive learning of Bayesian Networks using OpenMarkov", Disponible en http://www.openmarkov.org/learning/interactive_learning.pdf. Ultimo acceso: 9-2-2016.

[5] Agustín Cernuda, Mª Carmen Suárez, Daniel Gayo, Sonia Hevia, "Un estudio sobre el absentismo y el abandono en asignaturas de programación", *ReVisión*, 6 (1), 2013.

[6] Gregory F. Cooper, Edward Herskovits, "A Bayesian method for the induction of probabilistic networks from data", *Machine Learning*, 9 (4), pp. 309–347, 1992.

[7] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "From data mining to knowledge discovery in databases", *AI Magazine*, 17 (3), pp. 37, 1996.

[8] Elvira Consortium, "Elvira: An environment for probabilistic graphical models", Actas del First European Workshop on Probabilistic Graphical Models, pp. 222-230, 2002.

[9] Antonio Fernández, María Morales, Carmelo Rodríguez, Antonio Salmerón, "A system for relevance analysis of performance indicators in higher education using Bayesian networks", *Knowedge and Information Systems*, 27 (3), pp. 327–344, 2011.

[10] Alfonsa García, Ana Lías, Ángeles Mahíllo, Rosa Mª Pinero, "Abandono de primer año en la Ingeniería Informática", Actas de las JENUI 14, pp. 151-158, 2014.

[11] Horacio Kuna, Ramón García, Francisco R. Villatoro, "Identificación de causales de abandono de estudios universitarios. Uso de procesos de explotación de información", *Revista Iberoamericana de Tecnologías en Educación y Educación en Tecnología*, 5, pp. 39–44, 2009.

[12] Carmen Lacave, Manuel Luque, Francisco J. Díez. "Explanation of Bayesian networks and influence diagrams in Elvira", *IEEE Systems, Man and Cybernetics. Part B*, 37, pp. 952-965, 2007.

[13] Kin Fun Li, David Rusk, Fred Song, "Predicting Student Academic Performance", Actas del CISIS 2013, pp. 27–33, 2013.

[14] María Morales, Antonio Salmerón, "Análisis del alumnado de la Universidad de Almería mediante redes bayesianas", Actas del *27 Congreso Nacional de Estadística e Investigación Operativa*, 2003.

[15] Ajay Kumar Pal, Saurabh Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students", *International Journal of Electronics Communication and Computer Engineering*, 4 (5), pp 1560-1565, 2013.

[16] Judea Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[17] Joan Rué, "El abandono universitario: variables, marcos de referencia y políticas de calidad", *Revista de Docencia Universitaria,* 12(2), pp. 281-306, 2014.

[18] P. Spirtes, C. Glymour, R. Scheines, "Causation, prediction, and search", *Lecture Notes in Statistics*, 34, 1993.

[19] Kurt Thearling, "An Introduction to Data Mining: Discovering hidden value in your data warehouse", Disponible en http://www.thearling.com/text/dmwhite/dmwhite.htm. Ultimo acceso: 9-2-2016.