Minería de Datos en la Educación

Álvaro Jiménez Galindo
Hugo Álvarez García
Universidad Carlos III de Madrid
Avda. De la Universidad, 30
28911, Leganés (Madrid-España)
100055019@alumnos.uc3m.es
100064465@alumnos.uc3m.es

Abstract

En este documento se describe el uso de la minería de datos aplicada a entornos educativos y su uso pedagógico.

Categories and Subject Descriptors 1 [*Inteligencia Artificial*]: Minería de Datos.; 2 [*Educación*]: Entornos de aprendizaje, Pedagogía.

General Terms Teoría, aspectos educativos, relaciones, inferencia de información, mejora de programas educativos.

General Terms Minería de Datos, Educación, Pedagogía, Conjuntos de datos, Métodos de procesado, Modelos de inferencia.

1. INTRODUCCIÓN

La minería de datos, también conocida como Descubrimiento de Conocimiento en Bases de datos (sus siglas en inglés son "KDD – Knowledge Discovery in Databases"), es el campo que nos permite descubrir información nueva y potencialmente útil de grandes cantidades de datos. Se ha empleado en numerosos campos, incluyendo desde los ya conocidos casos de cesta de la compra hasta la bioinformática o investigaciones contra el terrorismo. Recientemente, se ha incrementado el interés en utilizar la minería de datos en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden. Los métodos empleados en la minería de datos en la educación suelen diferir de los métodos más generalistas, explotando explícitamente los múltiples niveles de jerarquía presentes en los datos. Métodos psicométricos suelen ser integrados con métodos de aprendizaje máquina y textos de minería de datos para lograr los objetivos. Por ejemplo, obteniendo datos sobre cómo los estudiantes eligen utilizar el software educacional, puede ser realmente útil considerar datos a distintos niveles sobre las pulsaciones de teclas, nivel de respuestas, del alumno, de la clase o de la escuela entera. Otros temas como el tiempo, secuencia o incluso el contexto juegan papeles importantes en el estudio de datos educacionales.

2. VENTAJAS RESPECTO A LOS PARADIGMAS TRADICIONALES DE INVESTIGACIÓN EDUCACIONAL

La minería de datos educacionales ofrece numerosas ventajas comparándola con los paradigmas más tradicionales de investigación relativa a la educación, como experimentos de laboratorio, estudios sociológicos o investigación de diseño. En particular, la creación de repositorios públicos de datos educacionales ha creado una base que hace posible la minería de datos educacionales. En particular, los datos de estos repositorios son totalmente válidos (ya que son datos reales sobre el rendimiento y aprendizaje de estudiantes reales, en ambientes educacionales, tomados en tareas de aprendizaje), y cada vez más fácilmente accesibles para comenzar una investigación. Estos puntos permiten a los investigadores ahorrar mucho tiempo en tareas como la búsqueda de individuos (tales como escuelas, profesores y alumnos), organización de los estudios y recopilación de datos, ya que estos se encuentran directamente accesibles. Aunque el uso de datos previamente recogidos limita los análisis a las cuestiones que conciernen a estos datos, una investigación previa puede resultar extremadamente útil para analizar cuestiones poco relacionadas con los datos tomados, como por ejemplo atributos de los estudiantes tales como comportamiento estratégico o motivación. La disponibilidad de estos datos ha supuesto un gran avance. Una vez definido un modelo de interés educativo sobre los datos, puede probarse con nuevos conjuntos de datos. La transferencia de estos modelos puede no ser trivial, pero el proceso de desarrollo y validación de un modelo para un nuevo contexto es mucho más rápido. Gracias a esta faceta, muchos análisis se han podido repetir sobre distintos sistemas o contextos de aprendizaje. Además, la existencia de miles de alumnos que usan herramientas de aprendizaje similares, aunque sea en distintos contextos, aporta una posibilidad nueva de estudiar la influencia de factores contextuales en profesores y alumnos. Históricamente, ha sido muy difícil estudiar cómo las diferencias entre grupos de profesores o clases influencian en aspectos específicos del aprendizaje. Este tipo de análisis resulta mucho más fácil con la minería de datos. De manera similar, el impacto de diferencias individuales ha sido difícil de estudiar estadísticamente con métodos tradicionales. La minería de datos aplicada al ambiente educativo posee el potencial de extender un conjunto de herramientas mucho más amplio para el análisis de cuestiones importantes sobre diferencias individuales.

3. PRINCIPALES ENFOQUES

Hay una gran variedad de métodos empleados habitualmente en el ámbito de la educación en la minería de datos. Estos métodos están comprendidos en las siguientes categorías: predicción, agrupamiento, minería de relaciones, inferencia a través de modelos, y destilación de datos para la interpretación por parte de un ser humano. Las tres primeras categorías son universales para distintos tipos de minería de datos (aunque en algunos casos con distintos nombres). Las categorías cuarta y quinta consiguen una particular importancia dentro de la minería de datos educacionales.

Cuadro 1. Principales enfoques de la minería de datos educacionales

Categoría del método	Objetivo del método	Aplicaciones Clave
Predicción	Desarrollo de un modelo que pueda inferir una variable a partir de la combinación de los datos disponibles	Detección de comportamiento del estudiante (engaños al sistema, distracciones, 'slipping'); Desarrollo de modelos de dominio; Predicción y entendimiento de los resultados académicos de un estudiante
Agrupamiento	Encontrar conjuntos de datos que se agrupen naturalmente, separando el conjunto completo en una serie de categorías	Descubrimiento de nuevos patrones de comportamiento de estudiantes; Investigación de similitudes y diferencias entre escuelas
Minería de relaciones	Descubrimiento de relaciones entre variables	Descubrimiento de asociaciones curriculares en secuencias de cursos; Descubrimiento de estrategias pedagógicas que guíen en un proceso más efectivo de aprendizaje
Descubrimiento mediante modelos	Modelado de un fenómeno mediante predicción, agrupamiento o ingeniería del conocimiento, es usado como componente en una futura predicción o minería de relaciones	Descubrimiento de relaciones entre comportamiento de estudiantes y sus características o variables contextuales; Análisis de cuestiones de investigación para una amplia variedad de contextos
Destilado de datos	Los datos son destilados para permitir a un humano identificar o clasificar rápidamente propiedades de los datos	Identificación humana de patrones en el aprendizaje de los alumnos, comportamiento colaboración; Etiquetado de datos para su uso en desarrollos posteriores de modelos predictivos

3.1. Predicción

En predicción, el objetivo es desarrollar un modelo que pueda inferir una variable a partir de alguna combinación de otras variables incluidas en los datos. La predicción requiere etiquetas para la variable de salida para un conjunto de datos limitado, donde una etiqueta suponga una información fiable sobre el valor de la variable de salida en casos específicos. De todas maneras, en algunos casos es importante considerar el grado en el que estas etiquetas puedan ser aproximadas o inciertas. La predicción tiene dos usos clave comprendidos en la minería de datos educacionales. En algunos casos, métodos de predicción pueden ser usados para estudiar qué características de un modelo son importantes para una predicción, dando información sobre la construcción subyacente. Este es un enfoque común en programas de investigación que tratan de predecir resultados educacionales sin predecir anteriormente factores intermedios. En un segundo tipo de uso, los métodos de predicción son utilizados para predecir cuál será el valor de salida en contextos donde no es deseable obtener una etiqueta para esa construcción (por ejemplo, en ocasiones en las que no haya datos etiquetados).

Como ejemplo, considérese una investigación que estudie la relación entre aprendizaje y engaño al sistema (en inglés, este término se refiere como 'game the system', definiéndose como el éxito en una tarea educativa tomando ventaja de la propiedades o regularidades del sistema usado para realizar dicha tarea, en vez de pensar y aprender a partir del material dado). Si un investigador tiene como objetivo estudiar esta construcción a través del uso de una herramienta de software durante un año completo en varias escuelas, puede no ser manejable el evaluar directamente, sin usar métodos de minería de datos, si un alumno está jugando con el sistema en cualquier punto y en cualquier momento. Baker et al desarrollaron un modelo de predicción usando datos recopilados automáticamente de interacciones entre estudiantes y el software como variables de predicción, y después validando la precisión del modelo al ser generalizado a más estudiantes y contextos. Entonces fueron capaces de estudiar sus avances en el conjunto completo de datos.

En general, existen tres tipos de predicción: clasificación, regresión y estimación de densidad. En clasificación, el valor predicho es una variable categórica o binaria. Algunos métodos populares de clasificación incluyen árboles de decisión, regresión logística (modelo de regresión para variables dependientes o de respuestas binomialmente distribuidas) y máquinas de soporte vector. En regresión, el valor predicho es una variable continua. Algunos métodos populares de regresión en la minería de datos educacionales incluyen la regresión lineal, redes neuronales y regresión sobre máquina de soporte vector. En estimación de densidad, la variable predicha es una función de densidad de probabilidad. Estimadores de densidad pueden estar basados en una variedad de funciones de kernel, incluyendo funciones gaussianas. Para cada tipo de predicción, las

variables de entrada pueden ser categóricas o continuas. Distintos métodos de predicción son más efectivos dependiendo en el tipo de variables de entrada utilizadas. Métodos populares para evaluar la precisión de una predicción incluyen la correlación lineal, coeficiente de Cohen Kappa (que tiene en cuenta los aciertos que se puedan producir por casualidad) y A' (el área bajo la curva 'receiver operating curve', una representación gráfica de la sensitividad). La precisión porcentual no suele ser usada para la clasificación, ya que es altamente dependiente de tasas de distintas clases. Entonces, una precisión muy alta puede ser lograda en algunos casos por un clasificador que siempre escoja la clase mayoritaria. Cuando se calcula la calidad de una predicción, es importante tener en cuenta la no-independencia de distintas observaciones que impliquen al mismo alumno. Para lograr este objetivo, para la minería de datos educacionales suelen aplicarse métodos meta-analíticos como Strube's Adjusted Z, o seleccionar estimadores conservadores que asuman completa dependencia.

3.2. Agrupamiento

En agrupamiento, el objetivo es encontrar puntos de datos que se agrupen de manera natural, repartiendo el conjunto original de datos en un conjunto de 'clusters'. El agrupamiento es particularmente útil en casos donde las categorías más comunes de los datos no son conocidas. Si un conjunto de clusters es óptimo, en cada categoría, cada punto será más similar a los puntos pertenecientes a su cluster que a puntos pertenecientes a otros grupos. Los clusters pueden ser creados con distinta granularidad: por ejemplo, las escuelas podrían ser agrupadas para investigar similitudes y diferencias entre ellas, los estudiantes podrían también agruparse por el mismo motivo, o incluso podrían agruparse las acciones de los estudiantes para investigar patrones de comportamiento. Los algoritmos de agrupamiento pueden comenzar sin hipótesis previas sobre los grupos de datos (como el algoritmo k-means con inicio aleatorio), o empezar desde una hipótesis específica, posiblemente generada en estudios previos con un conjunto de datos distinto. Un algoritmo de agrupamiento puede postular que cada punto debe pertenecer únicamente a un cluster (como en el algoritmo k-means), o puede decidir que algunos de los puntos pertenezcan a varios o ningún cluster (como en los modelos de mezcla de gaussianas). La calidad de un conjunto de grupos o clusters suele ser evaluada tomando como referencia la medida en la cual el conjunto de clusters se ajusta a los datos, relativo a cuánto se espera que se ajusten únicamente por casualidad dado el número de clusters, usando métricas estadísticas tales como el criterio de información bayesiano.

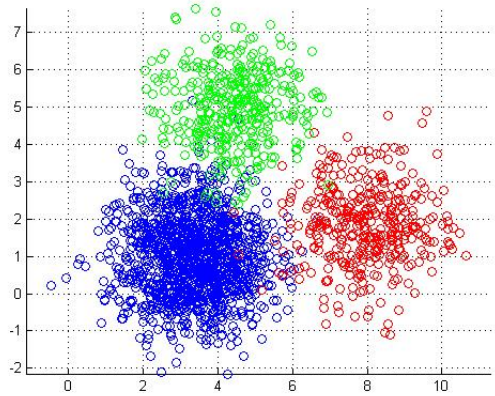

Figura 1. Conjunto Bayesiano de Clusters

3.3. Minería de Relaciones

En la minería de relaciones, el objetivo es descubrir relaciones entre variables en un conjunto de datos con un gran número de variables. Una forma de realizar esto es localizando las variables más fuertemente relacionadas con una única variable de interés, o también mediante el descubrimiento de las relaciones más fuertes entre dos variables. De manera general, existen cuatro tipos de minería de relaciones: minería de reglas de asociación, minería de correlación, minería de patrones de secuencias y minería de datos causales. En la minería de reglas de asociación, el objetivo es encontrar reglas 'si X entonces Y' de manera que si se encuentra un conjunto de variables, otra variable tendrá habitualmente un determinado valor. Por ejemplo, una regla podría ser:

$$\left\{ \begin{array}{c} \text{El estudiante} \\ \text{está frustrado} \\ \\ \text{El estudiante tiene} \\ \text{un sentimiento mas} \\ \text{fuerte de aprendizaje} \\ \text{que de éxito} \end{array} \right\} \Longrightarrow \left\{ \begin{array}{c} \text{El estudiante} \\ \text{habitualmente} \\ \text{pide ayuda} \end{array} \right\}$$

En la minería de correlaciones, el objetivo es encontrar correlaciones lineales (positivas o negativas) entre variables. En la minería de patrones de secuencias, el objetivo es encontrar asociaciones temporales entre eventos. Por ejemplo, para determinar qué secuencia de comportamientos de un estudiante da lugar eventualmente a un interés por el aprendizaje. En la minería de datos causales, el objetivo es descubrir si un evento ha sido la causa de otro evento, ya sea analizando la covarianza de los dos eventos o usando información sobre cómo uno de los eventos fue provocado. Por ejemplo, si un evento pedagógico es aleatoriamente escogido usando experimentación automatizada, y normalmente conlleva un resultado positivo de aprendizaje, una relación causal puede ser inferida. Las relaciones encontradas a través de la minería de relaciones deben satisfacer dos criterios: relevancia estadística y un determinado nivel de interés. La relevancia estadística es habitualmente evaluada a través de test estadísticos estándar, tales como los F-test. Debido a que un gran número de test son realizados, es necesario un control para encontrar relaciones casuales. Un método para realizar esto es usar métodos o ajustes estadísticos post-hoc que controlen el número de test realizados, como el ajuste de Bonferroni. Este método puede incrementar la confianza sobre una relación individual, descartando la posibilidad de que sea una casualidad. Un método alternativo es la evaluación de la probabilidad total del patrón de resultados obtenidos, usando métodos Monte Carlo. Este método evalúa cómo de probable es que el patrón total surgiese por casualidades. El nivel de interés de cada hallazgo es evaluado para reducir el conjunto de reglas / correlaciones / relaciones causales comunicadas a la persona que realiza la investigación. En conjuntos muy grandes de datos, cientos de miles de relaciones significativas pueden ser encontradas. El nivel de interés mide qué hallazgos son los más distintivos y mejor respaldados por los datos, en algunos casos también tratando de podar resultados similares. Hay una amplia variedad de medidas del nivel de interés, incluyendo el soporte, confianza, convicción, alzado, resaltado, cobertura, correlación y coseno. Algunas investigaciones sugieren que el alzado y el coseno pueden ser especialmente relevantes en el ámbito de los datos educacionales.

3.4. Descubrimiento mediante Modelos

En el descubrimiento mediante modelos, se desarrolla un modelo mediante predicción, agrupamiento o, en algunos casos, ingeniería del conocimiento (usando métodos de razonamiento humano en

vez de métodos automatizados). Este modelo es entonces utilizado como un componente en otro análisis, como predicción o minería de datos. En el caso de predicción, las predicciones hechas por el modelo creado son usadas como variables de entrada en la predicción de una nueva variable. Por ejemplo, el análisis de estructuras complejas como el engaño al sistema en el aprendizaje online habitualmente han dependido en la evaluación de la probabilidad de que un estudiante conociese ya de antemano el temario impartido. Estas evaluaciones del conocimiento del alumno han dependido a su vez de modelos de componentes del aprendizaje en un dominio, normalmente expresados como un mapeo entre ejercicios en el software de aprendizaje. En el caso de la minería de relaciones, se estudian las relaciones entre las predicciones del modelo creado y variables adicionales. Esto permite al investigador estudiar la relación entre una construcción compleja oculta y construcciones observables. A menudo, el descubrimiento mediante modelos enfatiza la validación generalizada de un modelo de predicción a través de varios contextos. Por ejemplo, Baker usó predicciones de engaños al sistema sobre datos de un año completo de software educacional para estudiar si factores de estado o características eran mejores predictores sobre cuánto un estudiante engañaría al sistema. La generalización se sustenta en una validación apropiada de que el modelo se comporta de la misma manera sobre varios contextos.

3.5. Destilado de datos

Otro área de interés en la minería de datos educacionales es la destilación de datos para la interpretación humana. En algunos casos, los seres humanos pueden realizar inferencias sobre datos cuando éstos son presentados adecuadamente, que se encuentran más allá del punto de mira inmediato de los métodos de minería de datos totalmente automatizados. Los métodos en esta área de minería de datos educacionales son de información y métodos de visualización. De todos modos, las visualizaciones más utilizadas en el campo educacional suelen ser distintas a las utilizadas para la resolución de problemas de visualización de la información, debiéndose a la estructura específica y el significado embebido en esa estructura, habitualmente presente en datos educacionales. Los datos son destilados para la interpretación humana por dos motivos clave: identificación y clasificación. Cuando los datos son destilados para identificación, son mostrados de manera que un ser humano pueda identificar patrones conocidos que son, sin embargo, difíciles de expresar formalmente. Por ejemplo, una visualización clásica de la minería de datos educacionales es la curva de aprendizaje, que representa el número de oportunidades de practicar una habilidad en el eje X, y muestra el rendimiento (como el porcentaje de aciertos o tiempo tomado para responder) en el eje Y.

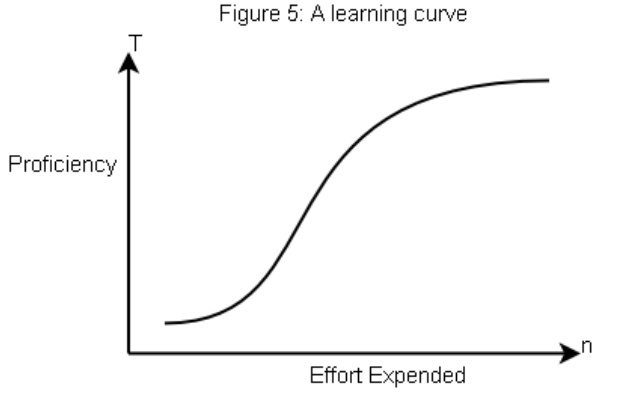

Figura 2. Curva de aprendizaje

Alternativamente, los datos pueden ser destilados para el etiquetado manual, dando soporte a desarrollos posteriores de un modelo predictivo. En este caso, sub-secciones de un conjunto de datos son mostrados en formato textual o visual, siendo etiquetados por codificadores humanos. Estas etiquetas son habitualmente utilizadas como base para el desarrollo de un predictor. Este enfoque ha demostrado acelerar en un factor de 40 los modelos predictivos de fenómenos complejos como los engaños al sistema, en relación con enfoques anteriores para la recolección de los datos necesarios.

4. APLICACIONES PRINCIPALES

Ha habido un amplio número de aplicaciones de minería de datos educacionales, como se ha podido ver a lo largo de este escrito. En esta sección, se ha prestado especial interés a cuatro áreas dentro del campo tratado. Un área clave de aplicación se encuentra en la mejora de los modelos de estudiante existentes, modelos que proporcionan información detallada sobre las características de un estudiante tales como el conocimiento, motivación, metaconocimiento y actitudes. El modelado de las diferencias individuales de cada estudiante para permitir al software responder a esas diferencias, es un tema clave en el desarrollo de software educativo. En los últimos años, la minería de datos educacionales ha permitido una expansión considerable en la sofisticación de los modelos de estudiantes. En particular, los métodos de minería de datos educacionales han permitido a los investigadores realizar inferencias de alto nivel acerca del comportamiento de los alumnos, tales como cuándo un estudiante está engañando al sistema, cuando un alumno se ha 'escurrido' (cometer un error a pesar de poseer la habilidad para responder la pregunta correctamente). Estos modelos de estudiante más avanzados han sido útiles en dos modos. En primer lugar, han incrementado nuestra habilidad para predecir el conocimiento de un alumno y su futuro rendimiento – la incorporación de modelos de acierto y de 'escurrirse' a las predicciones del rendimiento del estudiante han incrementado notablemente la exactitud de estas predicciones. En segundo lugar, estos modelos han permitido a los investigadores estudiar qué factores conducen al estudiante a tomar decisiones concretas en un entorno de aprendizaje. Una segunda área clave de aplicación es en el descubrimiento o mejora de modelos de la estructura de conocimiento del dominio. En la minería de datos educacionales, se han creado métodos para descubrir rápidamente modelos precisos directamente de los datos. Estos métodos han sido habitualmente combinados con marcos de modelado psicométricos con algoritmos avanzados de búsqueda en espacio, y son habitualmente planteados como problemas de predicción para el propósito de descubrimiento de modelos (por ejemplo, intentar predecir si acciones individuales serán correctas o incorrectas usando distintos modelos de dominio es un método común para el desarrollo de estos modelos). Un tercer área clave de la aplicación es el estudio del soporte pedagógico proporcionado por el software de aprendizaje. El software educacional moderno aporta distintos tipos de soporte pedagógico a los estudiantes. Descubrir cuál es el más efectivo ha sido un área de interés para los investigadores de minería de datos educacionales. La descomposición del aprendizaje, un tipo de minería de relaciones, ajusta datos de rendimiento a curvas de aprendizaje exponenciales, relacionando el éxito a la cantidad de cada tipo de soporte pedagógico que un estudiante ha recibido (con un peso para cada tipo de soporte). Los pesos indican cómo de efectivo es cada tipo de soporte pedagógico en la mejora del aprendizaje. Un ejemplo ilustrativo se ofrece en la siguiente sección. El cuarto área clave estudiado en la minería de datos educacionales son los descubrimientos científicos sobre el aprendizaje y los aprendices. Esto conlleva distintas formas. La aplicación de la minería de datos educacionales para la respuesta de preguntas

en cualquiera de las tres áreas anteriores pueden comprender beneficios científicos más amplios; por ejemplo, el estudio del soporte pedagógico puede tener un potencial futuro a largo plazo para enriquecer teorías sobre andamiaje. Más allá de estas áreas, sin embargo, ha habido muchos análisis enfocados directamente hacia el descubrimiento científico. El descubrimiento mediante modelos es un método clave para el descubrimiento científico a través de la minería de datos educacionales. La descomposición de métodos de aprendizaje es otro método prominente para llevar a cabo estudios científicos sobre el aprendizaje y los individuos implicados.

5. EJEMPLO DE ESTUDIO: Identificación de características de fracasos escolares en institutos

El abandono escolar siempre ha estado relacionado con factores sociales, económicos y psicológicos. Se ha intentado, a partir de ciertos estudios y usando distintas metodologías, identificar el proceso de un alumno con riesgo de fracaso escolar. En este caso, la minería de datos, junto con el uso de un modelo basado en árboles de decisión, nos ayudará a investigar las correlaciones existentes en los casos de fracaso escolar.

5.1. Estudios previos

Hess y Copeland, en 2001 ya midieron el uso de estrategias de copia por los estudiantes para construir un modelo de predicción con un análisis discriminante, ya que básicamente, era un proceso de clasificación personal. Se dieron cuenta de que el uso de ciertas estrategias de copia predecían con bastante acierto un fracaso en el instituto. De igual manera, Street y Franklin en 1991 se dieron cunenta que los estudiantes con un estatus socio-económico bajo, eran más propensos a abandonar prematuramente el instituto que los de un nivel socio-económico más elevado.

Pursley y Lan en 2003, elaboraron un excelente estudio sobre este tema, referenciando el abandono desde diferentes perspectivas, incluyendo los logros académicos, la motivación en el trabajo de la escuela, la participación en actividades, las aspiraciones educacionales, las percepciones de la escuela, las relaciones con los compañeros, y la autoestima.

Otro estudio reciente sobre las correlaciones existentes en este tema es el de Wayman en 2001. Se trata de un estudio muy técnico que toma muchas medidas de los estudiantes, a través de un modelo de regresión logística y de imputación múltiple[1]. Gracias a este estudio se econtró un conjunto muy potente de predictores basados en la recompensa del estudio, el nivel socio-económico, y la edad.

5.2. Árboles de decisión

Aplicando un modelo basado en árboles de decisión al sector de la educación, podremos identificar los estudiantes que requieran mayor ayuda en un área en particular. También, nos ayudarán a determinar el grupo de variables de predicción que estén más relacionadas con nuestra variable final, el índice de fracaso escolar. En el sector de los institutos de secundaria, las aplicaciones existentes de este tipo de estudios son escasas, y a menudo son proporcionadas únicamente en estudios "post-secundaria". En américa, es dificil encontrar un colegio K-12[2] que haya intentado un salto conceptual de lo que normalmente es conocido como un sofisticado proceso de negocio, a una aplicación de la minería de datos en la educiación pública.

La intención general del uso de árboles de decisión, como comentamos anteriormente, es encontrar el mejor predictor de la variable depediente ubicada en la raiz del propio árbol. Encontrar este predictor normalmente requiere recodificar o agrupar numerosos valores originales del predictor para crear al menos dos nodos. Cada nodo por tanto define la nueva rama del árbol que ha sido creada y para cada rama creada, el proceso vuelve a repetirse. El algoritmo busca el mejor predictor sobre el conjunto de variables restantes, y de nuevo, volverá a crear al menos dos ramas para ese mejor predictor. Cuando no se pueda encontrar un predictor que mejore la eficacia, el árbol no seguirá creciendo.

Vamos a discutir las ventajas que proporciona este tipo de modelado sobre los modelos estadísticos tradicionales. Primeramente, están diseñados para ser capaces de manejar un número muy grande de variables de predicción, en algunos casos, más allá de lo que permitiría el correspondiente modelo paramétrico estadístico. Otra ventaja, es que muchos modelos basados en árboles son completamente no-paramétricos y pueden capturar relaciones que los modelos paramétricos comunes no podrían manejar, o al menos no facilmente.

5.3. Análisis CHAID

"Chi-Squared Automatic Interaction Detection", en castellano "Detección Automática de Interacción basada en Chi-Cuadrado", es un método estadístico heurístico basado en arboles que examina las relaciones entre muchas variables de predicción categóricas, ordinales, o continuas, y la variable objeto de estudio. El programa empleado (Answer Tree, SPSS, 2001), proporciona un diagrama resumen (árbol), detallando las categorías que proporcionan mayor dependencia en nuestro objeto de estudio. También suministra una tabla para reportar qué nodos tienen la mayor concentración según un análisis de ganancias, y una tabla de información desclasificada según un análisis de riesgo.

5.4. Análisis del estudio

Estos análisis descritos en la sección anterior, a la vez sofisticados y elegantes, tienen una pega importante: es muy dificil explicarlo a personas sin conocimiento de estadística. Los individuos que posean poco o ningún entrenamiento de estadística, encuentran la regresión y sus otros primos paramétricos, unas metodologías un tanto desalentadoras para interpretarlas. Sin embargo, este sistema es sencillo y muy asequible para neófitos, ya que el análisis CHAID no abarca nada más complicado que un análisis frecuencial y de densidad, y el procedimiento de Chi-Cuadrado de Pearson es amigable y ampliamente conocido.

CHAID realiza comparaciones en pares para encontrar la variable de predicción más altamente relacionada con la variable raiz. En sistemas de muchas variables, tener esta función implementada en un ordenador es esencial para "picar" amplios conjuntos de datos.

Los datos empleados en este estudio fueron tomados de otros conjuntos de datos anteriores. No se empleó ninguna valoración adicional ni dato demográfico. Todas las variables empleadas fueron tomadas de bases de datos electrónicas del distrito. Los alumnos

[1] En la imputación múltiple, los valores que falten para estimar cualquier variable, se predicen usando valores existentes para otras variables. Estos valores predichos se llaman imputaciones, y constituyen el "conjunto de datos imputados".

[2] K-12 viene de "Kindergarden" (guardería, 4-6 años) hasta el 12º grado (16-19 años). Son el primer y último grado de educación gratuita en Estados Unidos, Australia, y la Canadá inglesa.

que estaban registrados como que habían abandonado el instituto durante un curso académico fueron comparados con una muestra aleatoria de alumnos que habían permanecido estudiando. A pesar de que alguna investigación sobre fracasos escolares, coomo la de Barrington y Hendricks, en 1989, encuentran poca relación entre el abandono y el sexo del alumno, también se incluyó en este experimento una variable que identificara el sexo del estudiante. A continuación mostramos una lista de las variables empleadas:

- Grupo (Abandono escolar/Graduado escolar)
- Edad en años
- Sexo
- Grupo étnico
- Status socio-económico
- Nº de infracciones disciplinarias de nivel 1
- Nº de infracciones disciplinarias de nivel 2
- Nº de infracciones disciplinarias de nivel 3
- Nº de infracciones disciplinarias de nivel 4
- Nº de clases avanzadas a las que ha asistido
- Nº de clases de Matemáticas a las que ha asistido
- Nº de clases de Ciencias a las que ha asistido
- Faltas justificadas
- Faltas sin justificar
- Nota media
- Nivel "CSAP Reading Proficiency" (entre cuatro niveles)
- Nivel "CSAP Writing Proficiency" (entre cuatro niveles)
- Nivel "CSAP Math Proficiency" (entre cuatro niveles)

Algunas de estas variables tuvieron q ser estimadas a partir de otros datos. Por ejemplo, el nivel socio-económico, fue estimado según si el colegio era gratuito o no, y según la posibilidad de elegir o no un almuerzo reducido. Estas relaciones a nuestro parecer pueden ser sólidas en Estados Unidos, pero quizás no tanto en paises europeos. El nivel de infraccion disciplinaria va de menor a mayor, siendo 1 el primero, y 4 el último. La nota media está escalada en 4 puntos, y los resultados del CSAP (Colorado Student Assessment Program) están medidos también en una escala del 1 al 4 (1:Insatisfactorio, 2:Parcialmente competente, 3:Competente, 4:Avanzado).

Como resultado de una ejecución preliminar usando todas estas variables, algunas de ellas fueron recodificadas con el fin de reducir el numero de categorías y hacer los árboles resultantes más interpretables. Se recodificaron "edad", "grupo étnico" y "nota media" y el número de clases de matemáticas, ciencias, y avanzadas. También, como el nivel "1:Insatisfactorio" y "4:Avanzado" ocurrían relativamente poco, se simplificaron las variables del CSAP con valores dicotómicos (1:Aprobado, 0:Suspenso).

Al final, sin embargo, los mejores predictores fueron las variables en sus estados sin recodificar.

5.5. Resultados

La siguiente tabla es el resultado de la construcción del arbol de decisión que incluía todas las variables originales como predictores potenciales. A los resultados, el software aplicó una validación cruzada con 25 muestras aleatorias. Como se puede ver, muestra la clasificación errónea de este modelo basado en árbol. La celda crítica (vista como un fracaso clasificado como un graduado), se mantiene en un mínimo (sólo 65 estudiantes) con este modelo.

Cuadro 2. Matriz de clasificación errónea

		Categoría Actual		
		Fracaso	Graduado	Total
Categoria Estimada	Fracaso	562	116	678
	Graduado	65	403	468
	Total	627	519	1146
		Riesgo Estimado 0.158		Validación Cruzada 0.172

Quizás el inconveniente del modelo es que el riesgo estimado (verosimiltud de todos los tipos de clasificación errónea), está entorno al 16 %. Por otra parte, la predicción debe ser correcta el 80 % de las veces.

Al inpeccionar el árbol generado[3], se observa que la variable más relacionada con el fracaso escolar, es el rendimiento académico. La primera variable (con mayor estadística chi-cuadrado) es la nota media. Los nodos 1 a 6 correspondientes al primer nivel por debajo de la raiz, realizan por tanto la primera clasificación en el árbol, siendo el nodo 1 los que reciben la nota más baja, y nodo 6 los de la nota más alta. Según estos nodos, los patrones de clasificación van cambiando. Por ejemplo, para los de la nota más baja, se les clasifica según la edad, de tal manera que si tiene mala nota y mucha edad tiene más probabilidad de fracaso que si tiene mala nota pero es más joven. A pesar de que el nodo 1 sea dividido, las probabilidades de fracaso son muy altas y por tanto se podría simplificar aún más el árbol eliminando estos subnodos.

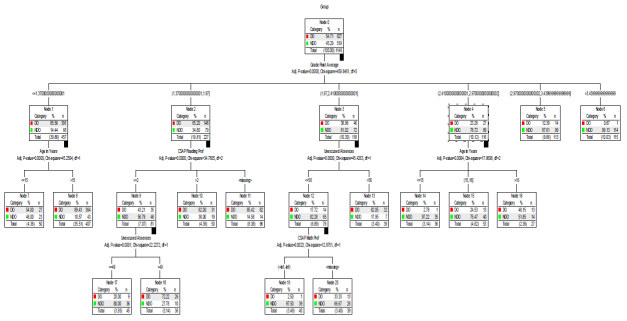

Figura 3. Árbol completo

El nodo 2 (correspondiente a los q tienen una nota media entre 1,37 y 1,97 sobre 4), se divide según la evaluación de la prueba CSAP Reading. Esto nos está indicando que los alumnos tienen mayor probabilidad de permanecer en el colegio si se han presentado a este test, y más aún si tienen habilidades de lectura. También se extrae que aquellos con más de 56 horas de faltas sin justificar tienen mayor probabilidad de fracaso. El nodo 4 (con nota media entre 2.41 y 2.97) se divide acorde la edad. Los nodos 5 y 6 no se dividen, ya que las variables en el patrón subyacente a notas medias altas no tienen valor potencial en la estimación de nuestro objetivo.

En un tercer nivel, bajo el nodo 12 (nota media entre 1.97 y 2.41 y además menos de 56 horas de faltas sin justificar) se subdivide según se han presentado o no a la prueba de matemáticas. De los

[3] Hemos incluido el mismo árbol en un apéndice para verlo más claramente

que se encuentran en esta situación y se han presentado a la prueba de matemáticas (independientemente de si se han presentado o no), sólo el 2,5 % de los alumnos abandonan la escuela, lo cual nos resulta bastante sorprendente, ya que tienen menos probabilidad de abandono que los que tienen una nota media de 2.97 a 3.44.

5.6. Conclusiones del estudio

Como empezamos explicando al comienzo del caso de estudio, el propósito de este se podía dividir en dos: primero investigar la existencia de variables relacionadas con el fracaso escolar, y segundo, aplicar minería de datos sobre las fuentes existentes con árboles de decisión. El arbol obtenido proporciona cierta habilidad para predecir qué estudiantes tienen riesgo de fracasar.

El programa empleado fue Answer Tree. Se trata de un programa muy completo y útil para elaborar estos árboles, y además tiene características adicionales accesibles de manera interactiva sobre le arbol obtenido, que no explicamos en este trabajo para no desviarnos en exceso del tópico central.

Aplicando este sistema, con la debida instrucción al personal del instituto (no excesiva, ya que no tienen por qué comprender los principios de funcionamiento del sistema) y gracias a la minería de datos, se puede hacer un seguimiento más cercano a los estudiantes que tengan mayor riesgo de abandono. Además, sería facil generar bases de datos y seleccionar grupos de trabajo para los alumnos que tengan carencias educacionales específicas y éstas les incremente peligrosamente el riesgo de fracaso escolar.

Este sistema debería actualizarse anualmente, y más ahora que los medios interactivos están cambiando la enseñanza en las escuelas y estos podrían insertar nuevas variables en nuestro árbol de decisión. Se podrá observar cómo estos nuevos métodos educacionales (pizarras electrónicas, ordenadores portátiles, portales académicos como moodle..) cambian nuestro árbol de decisión y actuar en función de las nuevas variables involucradas en nuestro árbol para lograr maximizar nuestro objetivo, el éxito académico.

Referencias

[1] William R. Veitch, Ph.D. Identifying Characteristics of High School Dropouts: Data Mining With A Decision Tree Model. Presentado en el *Annual Meeting of the American Educational Research Association* San Diego, CA April, 2004

[2] Baker, R.S.J.d. Data Mining for Education. Encontrado en McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*. Oxford, UK: Elsevier

[3] Cristóbal Romero, Sebastián Ventura, Enrique García. *Data mining in course management systems: Moodle case study and tutorial.* Department of Computer Sciences and Numerical Analisys, University of Córdoba, 14071 Córdoba, Spain

[4] Enrique García, Cristóbal Romero, Carlos de Castro, Sebastián Ventura. *Usando minería de datos para la contínua mejora de cursos de E-Learning.* Presentado en *Conferencia IADIS Ibero-Americana WWW/Internet 2006.* Escuela Politécnica Superior. Universidad de Córdoba.

[5] Antonio González-Pardo, Francisco B. Rodríguez, Estrella Pulido and David Camacho. *Using Virtual Worlds for Behaviour Clustering-based Analysis.* Departamento de Ingeniería Informática. Escuela Politécnica Superior. Universidad Autónoma de Madrid.

A. Apéndice: Árbol Completo

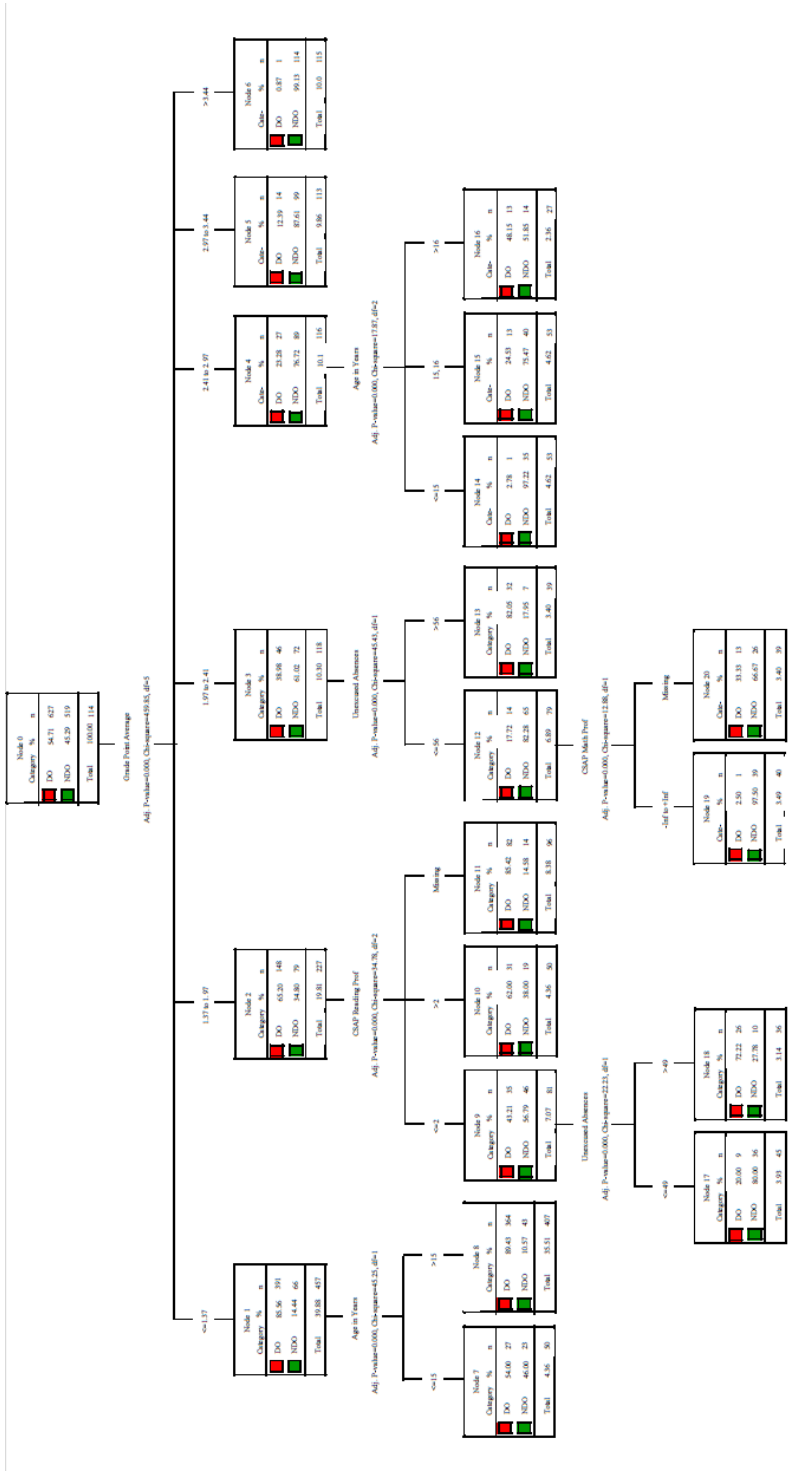

Figura 4. Árbol Completo