Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos

Carlos Márquez Vera[1], Cristóbal Romero Morales[2] y Sebastián Ventura Soto[3].

***Title*- Predicting of school failure using data mining techniques.**

***Abstract*- This paper proposes to apply data mining techniques to predict school failure and drop out. We use real data on 670 middle-school students from Zacatecas, México and employ white-box classification methods such as induction rules and decision trees. Experiments attempt to improve their accuracy for predicting which students might fail or drop out by: firstly, using all the available attributes; next, selecting the best attributes; and finally, rebalancing data, and using cost sensitive classification. The outcomes have been compared and the best resulting models are shown.**

***Index Terms*— School failure, Drop out, Educational Data Mining, Prediction, Classification.**

I. Introducción

En los últimos años ha surgido en muchos países una preocupación ante el problema del fracaso escolar y un creciente interés por determinar los múltiples factores que pueden influir en él [1]. La mayoría de los trabajos que intentan resolver este problema [2] están enfocados en determinar cuáles son los factores que más afectan al rendimiento de los estudiantes (abandono y fracaso) en los diferentes niveles educativos (educación básica, media y superior) mediante la utilización de la gran cantidad de información que los actuales equipos informáticos permiten almacenar en bases de datos. Todos estos datos constituyen una auténtica mina de oro de valiosa información sobre los estudiantes. El problema es que identificar y encontrar información útil y oculta en grandes bases de datos es una tarea difícil [3]. Una solución muy prometedora para alcanzar este objetivo es el uso de técnicas de extracción de conocimiento o minería de datos en educación, lo que ha dado lugar a la denominada minería de datos educativa (*Educational Data Mining*, *EDM*) [4]. Esta nueva área de investigación se ocupa del desarrollo de métodos para explorar los datos que se dan en el ámbito educativo, así como de la utilización de estos métodos para entender mejor a los estudiantes y los contextos en que ellos aprenden [5]. Las técnicas de EDM ya se han empleado con éxito para crear modelos de predicción del rendimiento de los estudiantes [6], obteniendo resultados prometedores que demuestran cómo determinadas características sociológicas, económicas y educativas de los alumnos pueden afectar en el rendimiento académico [7]. Es importante también destacar que hasta la fecha la mayor parte de las investigaciones sobre minería de datos aplicada a los problemas de abandono y fracaso, se han aplicado, sobre todo, en el nivel de educación superior [8] y, en mayor medida en la modalidad de educación a distancia [9]. Por el contrario, se ha encontrado muy poca información sobre la aplicación en la educación básica o media, donde sólo se han realizado simples análisis de la información basados en métodos estadísticos [10]. Existen algunas diferencias y/o ventajas entre aplicar minería de datos con respecto a sólo utilizar modelos estadísticos [11]:

- La minería de datos es más amplia ya que es un proceso completo formado por varias etapas y que incluye muchas técnicas, entre ellas, las estadísticas. Este proceso de descubrimiento de información está formado por las etapas de pre-procesado, la aplicación de técnicas de minería de datos (una de ellas puede ser estadística) y la evaluación e interpretación de los resultados.
- En las técnicas estadísticas (análisis de datos) se suele utilizar como criterio de calidad la verosimilitud de los datos dado el modelo. En minería de datos suele utilizar un criterio más directo, por ejemplo, utilizando el porcentaje de datos bien clasificados.
- En estadística la búsqueda se suele realizar mediante modelización basada en un algoritmo de ascenso de colinas (hill-climbing) en combinación con un test de hipótesis basado en razón de verosimilitud. En minería de datos se suele utilizar una búsqueda basada en meta-heurísticas.
- La minería de datos está orientada a trabajar con cantidades muy grandes de datos (millones y billones de datos). En cambio la estadística no suele funcionar tan bien en bases de datos de tan gran tamaño y alta dimensionalidad.

En este trabajo se propone la utilización de técnicas de minería de datos para detectar, cuáles son los factores que más influyen para que los estudiantes de enseñanza media o secundaria fracasen, es decir, suspendan o abandonen. Además se propone utilizar diferentes técnicas de minería de datos debido a que es un problema complejo, los datos suelen presentar una alta dimensionalidad (hay muchos factores que pueden influir) y suelen estar muy desbalanceados (la mayoría de los alumnos suelen aprobar y sólo una minoría suele fracasar). El objetivo final es detectar lo antes posible a los estudiantes que presenten esos factores para poder ofrecerles algún tipo de atención o ayuda para tratar de evitar y/o disminuir el fracaso escolar.

[1] Carlos Márquez Vera, es profesor de la Universidad Autónoma de Zacatecas, México. Jardín Juárez 147, 98000 (Teléfono: 4929229471, e-mail:z92mavec@uco.es)

[2] Cristóbal Romero Morales es profesor de la Universidad de Córdoba, Campus de Rabanales, 14071 Córdoba, España (Teléfono: 34-957-212172; fax: 34-957-218630; email: in1romoc@uco.es).

[3] Sebastián Ventura Soto es profesor de la Universidad de Córdoba, Campus de Rabanales, 14071 Córdoba, España (sventura@uco.es).

DOI (Digital Object Identifier) Pendiente

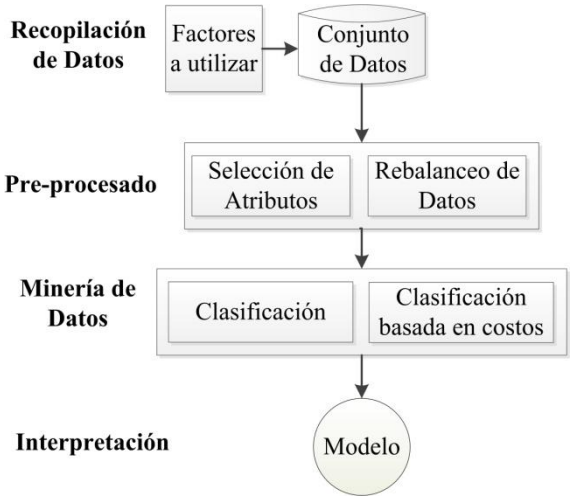

Fig. 1. Método utilizado para predicción del fracaso escolar.

Este artículo está organizado de la siguiente forma: En la siguiente sección se hace una descripción general del método que proponemos para predecir el fracaso escolar. En la sección III, se describen los datos que hemos utilizado para este trabajo y las fuentes de información de las cuales fueron extraídos. La sección IV explica las tareas de pre-procesado de los datos que hemos realizado. La sección V describe las diferentes pruebas de minería de datos que hemos realizado y los resultados obtenidos de las mismas. En la sección VI se hace una interpretación de los resultados y finalmente en la sección VII, se describen las conclusiones del trabajo y las futuras líneas de investigación.

II. Método

El método que se propone en este artículo está basado en la utilización de técnicas de minería de datos (ver Figura 1) y se compone de los pasos típicos de un proceso de extracción de conocimiento.

- Recopilación de datos. En esta etapa se recoge toda la información disponible de los estudiantes. Para ello primero se debe de seleccionar el conjunto de factores que pueden afectar y después se deben de recoger a partir de las diferentes fuentes de datos disponibles. Finalmente toda esta información se debe integrar en un único conjunto de datos.
- Pre-procesado. En esta etapa se preparan los datos para poder aplicar, posteriormente, las técnicas de minería de datos. Para ello, primero se realizan tareas típicas de pre-procesado como: limpieza de datos, transformación de variables y particionado de datos. Además se han aplicado otras técnicas como la selección de atributos y el re-balanceado de datos para intentar solucionar los problemas de la alta dimensionalidad y desbalanceo que presentan normalmente este tipo de conjuntos de datos.
- Minería de datos. En esta etapa se aplican algoritmos de minería de datos para predecir el fracaso escolar como si fuera un problema de clasificación. Para ello, se propone utilizar algoritmos de clasificación basada en reglas y en árboles de decisión debido a que son técnicas de "caja blanca" que generan modelos altamente interpretables que permiten su utilización directa en procesos de toma de decisiones. Además de la clasificación tradicional se propone utilizar también clasificación basada en costos o penalizaciones para intentar corregir el problema del desbalanceo de los datos. Finalmente, los distintos algoritmos utilizados deben de ser evaluados y comparados para determinar cuáles obtienen los mejores resultados de clasificación.
- Interpretación de los resultados. En esta última etapa, se analizan los modelos que han obtenido los mejores resultados para utilizarlos en la detección del fracaso escolar. Para ello, se analizan los factores que aparecen en las reglas y/o árboles de decisión, los valores que presentan y como están relacionados con otros factores.

A continuación, se describe un caso de estudio realizado con datos obtenidos de alumnos reales para mostrar la utilidad del método propuesto.

III. Recopilación de Datos

El fracaso escolar es conocido como "el problema de las mil causas" [12] debido a la gran cantidad posible de causas o factores de tipo personal, académicas, físicas, económicas, familiares, sociales, institucionales, pedagógicas, etc., que pueden tener influencia en el fracaso o abandono de los estudiantes. En nuestro caso de estudio concreto, los datos que hemos utilizado son de estudiantes del Programa II de la Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México. Todos los estudiantes que han participado en este estudio eran de nuevo ingreso en el curso académico 2009-2010 en el nivel medio-superior de educación Mexicana. Toda la información se recopiló de tres fuentes diferentes:

a) De una encuesta que se les aplicó a todos los alumnos a mitad del curso, con la finalidad de obtener información para detectar factores importantes que pueden incidir en su rendimiento escolar.
b) Del Centro Nacional de Evaluación (CENEVAL). Organismo que, entre otras actividades, realiza exámenes de ingreso o admisión en muchas instituciones de educación media y superior. Cuando los estudiantes se inscriben al examen, también se les hace un estudio socioeconómico, de éste se extrajo parte de la información.
c) Del Departamento de Servicios Escolares del Programa II, donde se recogen todas las notas obtenidas por los estudiantes.

A continuación, en la Tabla I, se muestran todas las variables agrupadas en las tres fuentes de datos que hemos utilizado.

IV. Pre-procesado de Datos

Antes de aplicar un algoritmo de minería de datos, generalmente hay que realizar algunas tareas de pre-procesado, que permitan transformar los datos originales a una forma más adecuada para ser usada por el algoritmo en particular. En nuestro caso de estudio estas tareas han consistido en la integración, la limpieza, la transformación y la discretización de datos [13].

La integración de los datos consiste en agrupar toda la información disponible de cada estudiante de las tres fuentes de datos en un único fichero de datos:

TABLA I
VARIABLES UTILIZADAS Y FUENTE DE PROCEDENCIA

Fuente	Variable
Encuesta	Semestre y grupo, Turno, Nivel de motivación, Sanción administrativa, No. de amigos, Tiempo de estudio adicional, Forma de estudio, Lugar de estudio, Cuándo estudia, Dudas, E. Civil, Si tiene hijos, Religión, Carrera universitaria electa, Influencia en decisión de carrera universitaria, Personalidad, Discapacidad física, Enfermedad grave, Bebidas alcohólicas, Si fuman, Nivel económico, Recursos para estudiar, Beca, Trabajo, Con quién vive, Nivel educativo de la madre, Nivel educativo del padre, No. de hermanos, Orden de nacimiento, Espacio para estudiar, Estímulo de los padres, Habitantes en comunidad, Años viviendo en comunidad, Modo de transporte, Distancia a la escuela, Asistencia a clases, se aburre en clase, Considera los conocimientos útiles, Asignatura difícil, Toma apuntes, Exceso de tarea, No. De alumnos en grupo, Forma de enseñar, Infraestructura escolar, Asesor, Interés de la institución.
CENEVAL	Edad, Sexo, Estado de procedencia, Régimen de escuela de procedencia, Modelo de secundaria, Promedio de secundaria, Trabajo de la madre, Trabajo del padre, No. de PC en familia, Limitado para ejercicio, Frecuencia de ejercicio, Tiempo de sesiones de ejercicio, Nota en Razonamiento Lógico Matemático, Nota en Matemáticas, Nota en Razonamiento Verbal, Nota en Español, Nota en Biología, Nota en Física, Nota en Química, Nota en Historia, Nota en Geografía, Nota en Formación Cívica, Nota en Ética, Nota en Inglés y calificación del EXANI I.
Departamento Escolar	Nota en Matemáticas 1, Nota en Física 1, Nota en Ciencias Sociales 1, Nota en Humanidades 1, Nota en Taller de Lectura y Redacción 1, Nota en Inglés 1, Nota en Computación 1, Estado Académico.

a) Los datos de la encuesta en papel se pasaron a formato electrónico.
b) Los datos del CENEVAL ya se encontraban en formato electrónico.
c) Los datos del departamento escolar también se encontraban en formato electrónico.

En la etapa de limpieza, se extrajo del conjunto de datos a aquellos estudiantes que no tenían completa al 100% toda la información. Es decir, que si durante el estudio socioeconómico que realizó el CENEVAL o durante la encuesta para detectar los factores que afectan el desempeño académico, algún estudiante omitió una o más respuestas, entonces éste se excluyó del conjunto de datos.

En la etapa de transformación se modificaron algunos nombres de atributos e instancias que contenían la letra Ñ, porque el software de minería de datos utilizado cambia este caracter por otro símbolo poco usual. Además también se puso la edad en años de los estudiantes, debido a que la información proporcionada por el CENEVAL, contenía, día, mes y año de nacimiento.

En la etapa de discretización, tanto las calificaciones o notas obtenidas en el examen de admisión EXANI I, el promedio en la secundaria y en las asignaturas cursadas durante el semestre, cambiaron de formato numérico (valores de 0 a 10) a formato nominal o categórico. Concretamente las etiquetas utilizadas y los rangos de discretización de las notas fueron: Excelente (9,5 a 10); Muy Bien (8,5 a 9,4); Bien (7,5 a 8,4); Regular (6,5 a 7,4); Suficiente (6,0 a 6,4); Deficiente (4,0 a 5,9); Muy Deficiente (menor a 4,0) y No presentó. Se creó un fichero con formato .ARFF de Weka [14], que es el software de minería de datos elegido para realizar las pruebas.

Después de realizar las anteriores tareas de pre-procesado, se dispone de un primer fichero de datos con 77 atributos/variables sobre 670 alumnos. Este fichero de datos fue particionado (10 particiones) para poder hacer una validación cruzada en las pruebas de clasificación. Una partición es la división aleatoria del fichero original de datos en otros dos, uno para la etapa de entrenamiento (*training*) y el otro para la etapa de prueba (*test*).

Debido a la gran cantidad de atributos recopilados (77), se realizó también un análisis o estudio de selección de atributos para determinar cuáles son los que mayormente influyen en la variable de salida o clase a predecir (Estado Académico). Para seleccionar las variables de mayor relevancia se utilizaron varios métodos de selección de atributos disponibles en el software Weka. En general, estos algoritmos de selección pueden ser agrupados por varios criterios. Una categorización popular es aquella en la que los algoritmos se distinguen por su forma de evaluar atributos y se clasifican en: filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje y w*rappers* (envoltorios), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto [15].

A continuación, en la Tabla II, se muestran los resultados obtenidos de la aplicación de 10 algoritmos de selección de mejores atributos sobre el fichero de datos.

Para seleccionar los mejores atributos se revisaron los resultados obtenidos por los 10 algoritmos de selección y se contabilizaron los que han sido seleccionados por varios de ellos. En la Tabla III se muestra la frecuencia de aparición de cada atributo, de esta tabla seleccionamos solamente aquellos con frecuencia mayor de 2, es decir, los mejores atributos son los que al menos 2 algoritmos los han seleccionado.

Al hacer la selección de los atributos con mayor frecuencia, se ha pasado de tener los 77 atributos originales a solamente los 15 mejores. Nuevamente, este fichero de datos se partió en 10 ficheros de entrenamiento y 10 ficheros de prueba.

Finalmente, como ya se mencionó anteriormente, nuestro conjunto de datos está desbalanceado. Este problema ocurre cuando el número de instancias de una clase es mucho menor que el número de instancias de la otra clase (o de otras clases). En nuestro caso, de los 670 alumnos, 610 aprobaron y 60 suspendieron o abandonaron. Por lo que se considera que los datos están desbalanceados, es decir, hay una mayoría de alumnos que aprobaron frente a una minoría que suspendieron. El problema de utilizar datos desbalanceados es que los típicos algoritmos de clasificación han sido desarrollados para maximizar la tasa de exactitud total, lo cual es independiente de la distribución de clases, esto provoca que los clasificadores tengan en la etapa de entrenamiento una clase mayoritaria lo que lleva a clasificar en la etapa de prueba con baja sensibilidad a los elementos de la clase minoritaria. Una manera de resolver el problema, es actuar en la etapa de pre-procesado de los datos, haciendo un sobre muestreo o balanceo de la distribución de clases, para ello existen varios algoritmos de re-balanceo y uno ampliamente usado es el denominado SMOTE (Synthetic Minority Oversampling Technique) disponible en Weka como un filtro de datos.

TABLA II
MEJORES ATRIBUTOS SELECCIONADOS

Algoritmo	Atributos Seleccionados
CfsSubsetEval	Discapacidad física; Edad; Matemáticas 1; Física 1; Ciencias Sociales 1; Humanidades 1; Taller de Lectura y Redacción 1; Inglés 1; Computación 1.
ChiSquared-AttributeEval.	Humanidades 1; Inglés 1; Ciencias Sociales 1; Física 1; Matemáticas 1; Computación 1; Taller de Lectura y Redacción 1; Nivel de motivación.
Consistency-SubsetEval	Semestre y grupo; Sesiones de ejercicio; Humanidades 1; Inglés 1, Dudas.
Filtered-AttributeEval	Humanidades 1; Inglés 1; Matemáticas 1; Ciencias Sociales 1; Física 1; Taller de Lectura y Redacción 1; Computación 1; Nivel de motivación; Promedio de secundaria; Historia; Semestre y grupo; Calificación de EXANI I.
FilteredSubsetEval	Matemáticas 1; Física 1; Ciencias Sociales 1; Humanidades 1; Taller de Lectura y Redacción 1; Inglés 1; Computación 1.
GainRatio-AttributeEval	Matemáticas 1; Humanidades 1; Inglés 1; Ciencias Sociales 1; Física 1; Taller de Lectura y Redacción 1; Computación 1; Nivel de motivación; Estado civil; Discapacidad física; Promedio de secundaria; Fumas.
InfoGain-AttributeEval	Humanidades 1; Inglés 1; Matemáticas 1; Ciencias Sociales 1; Física 1; Taller de Lectura y Redacción 1; Computación 1.
OneRAttributeEval	Humanidades 1; Ciencias Sociales 1; Inglés 1; Computación 1; Taller de Lectura y Redacción 1; Matemáticas 1; Física 1; Nivel de motivación.
ReliefFAttributeEval	Física 1; Inglés 1; Matemáticas 1; Humanidades 1; Taller de Lectura y Redacción 1; Ciencias Sociales 1; Promedio de secundaria; Computación 1; Nivel de motivación; Edad; Calificación de EXANI I; Fumas.
SymmetricalUncert-AttributeEval	Humanidades 1; Matemáticas 1; Inglés 1; Ciencias Sociales 1; Física 1; Taller de Lectura y Redacción 1; Computación 1.

TABLA III
ATRIBUTOS DE MAYOR INFLUENCIA ORDENADOS SEGÚN LA FRECUENCIA DE APARICIÓN EN LOS MÉTODOS DE SELECCIÓN DE ATRIBUTOS

Atributo	Frecuencia
Humanidades 1	10
Inglés 1	10
Ciencias Sociales 1	9
Matemáticas 1	9
Taller de Lectura y Redacción 1	9
Física 1	9
Computación 1	9
Nivel de motivación	5
Promedio de secundaria	3
Fumas	2
Calificación de EXANI I	2
Edad	2
Discapacidad física	2
Semestre y grupo	2

En términos generales, SMOTE introduce de manera sintética elementos de la clase minoritaria para equilibrar la muestra de datos, basado en la regla del vecino más cercano. Los elementos sintéticos creados son introducidos en el espacio que hay entre los elementos de la clase minoritaria. Dependiendo del tamaño del sobre muestreo requerido los vecinos más cercanos son elegidos aleatoriamente [16]. En nuestro caso el conjunto de datos con los 15 mejores atributos y con 670 instancias fue particionado de la siguiente manera: cada fichero de entrenamiento se re-balanceó con el algoritmo SMOTE de forma que tuviera el 50% de instancias Aprobó y el 50% de instancias Suspendió, dejando los ficheros de prueba sin re-balancear.
Después de realizar todas las tareas de pre-procesado de datos, se cuenta con:

- 10 ficheros de entrenamiento y testeo con todos los atributos (77 atributos).
- 10 ficheros de entrenamiento y testeo con sólo los 15 mejores atributos.
- 10 ficheros de entrenamiento y testeo con sólo los 15 mejores atributos, donde los ficheros de entrenamiento están re-balanceados.

V. MINERÍA DE DATOS Y EXPERIMENTACIÓN

En esta sección se describen los experimentos realizados y las técnicas de minería de datos utilizadas para la obtención de modelos de predicción del Estado Académico de los estudiantes al final del semestre.

Concretamente, hemos realizado varios experimentos con el objetivo de obtener la máxima exactitud de clasificación. En un primer experimento hemos ejecutado 10 algoritmos de clasificación utilizando todos los atributos con los que se cuenta, es decir de toda la información disponible. En un segundo experimento, utilizamos sólo los mejores atributos o variables. En un tercer experimento, hemos repetido las ejecuciones pero utilizando los ficheros de datos re-balanceados. En un último experimento hemos consideramos diferentes costos de clasificación.

Hemos seleccionado 10 algoritmos de clasificación de entre los disponibles por la herramienta de minería de datos Weka. Esta selección se ha realizado debido a que estos algoritmos, son todos del tipo "caja blanca", es decir, se obtiene un modelo de salida comprensible para el usuario, porque o se obtienen reglas de clasificación del tipo "Si – Entonces" o árboles de decisión. De esta forma un usuario no experto en minería de datos como un profesor o instructor puede utilizar directamente la salida obtenida por estos algoritmos para detectar a los alumnos con problemas a tiempo y poder tomar decisiones sobre cómo ayudarlos y evitar que suspendan o abandonen.

Las reglas de clasificación del tipo "Si – Entonces" son una manera simple y fácilmente comprensible de representar el conocimiento. Una regla tiene dos partes, el antecedente y el consecuente. El antecedente de la regla (la parte del "Si") contiene una combinación de condiciones respecto a los atributos de predicción. El consecuente de la regla (la parte del "Entonces") contiene el valor predicho para la clase. De esta manera, una regla asigna una instancia de datos a la clase señalada por el consecuente si los valores de los atributos de predicción satisfacen las condiciones expresadas en el antecedente, y por tanto, un clasificador es representado como un conjunto de reglas. Los algoritmos incluidos en este paradigma pueden ser considerados como una búsqueda heurística en un espacio de estados. En este caso, un estado corresponde a una regla candidata, y los operadores corresponden a la generalización y especialización de operaciones que transformen una regla candidata en otra. Los 5 algoritmos de inducción de reglas de clasificación que se usaron son: JRip, NNge, OneR, Prism [17] y Ridor.

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, el cual contiene cero o más nodos internos y uno o más nodos de hoja. Los nodos internos tienen dos o más nodos secundarios y contienen divisiones, los cuales prueban el valor de una expresión de los atributos. Los arcos de un nodo interno a otro secundario (o de menor jerarquía) son etiquetados con distintas salidas de la prueba del nodo interno. Cada nodo hoja tiene una etiqueta de clase asociada. El árbol de decisión es un modelo predictivo en el cual una instancia es clasificada siguiendo el camino de condiciones cumplidas desde la raíz hasta llegar a una hoja, la cual corresponderá a una clase etiquetada. Un árbol de decisión se puede convertir fácilmente en un conjunto de reglas de clasificación [18]. Los 5 algoritmos de árboles de decisión que se utilizarán son J48 [18], SimpleCart [19], ADTree [20], RandomTree y REPTree.

En el primer experimento, se han ejecutado los 10 algoritmos utilizando toda la información disponible, es decir, los ficheros de datos con los 77 atributos de los 670 alumnos. Hemos realizado una validación cruzada con 10 particiones. En este tipo de validación cruzada, se realiza el entrenamiento y el testeo diez veces con las diferentes particiones. Los resultados obtenidos (la media de las 10 ejecuciones) con los ficheros de prueba/test de la aplicación de los algoritmos de clasificación se muestran en la Tabla IV. Se ha indicado además del porcentaje de exactitud global o total, también los porcentajes para cada uno de los dos valores de la clase (Aprobó y Suspendió/Abandonó) y una medida de tendencia central bastante empleada para casos similares a éste de datos desbalanceados, la media geométrica.

TABLA IV
VALIDACIÓN CRUZADA UTILIZANDO LOS 77 ATRIBUTOS DISPONIBLES

Algoritmo	%Aciertos Aprobó	%Aciertos Suspendió	%Exactitud Total	Media Geométrica
JRip	97,7	**78,3**	96,0	**87,5**
NNge	98,5	73,3	96,3	85,0
OneR	98,9	41,7	93,7	64,2
Prism	99,5	25,0	93,1	49,9
Ridor	96,6	65,0	93,7	79,2
ADTree	**99,7**	76,7	**97,6**	87,4
J48	97,4	53,3	93,4	72,1
RandomTree	95,7	48,3	91,5	68,0
REPTree	98,0	56,7	94,3	74,5
SimpleCart	97,7	65,0	94,8	79,7

TABLA V
VALIDACIÓN CRUZADA UTILIZANDO LOS 15 ATRIBUTOS SELECCIONADOS COMO MEJORES

Algoritmo	%Aciertos Aprobó	%Aciertos Suspendió	%Exactitud Total	Media Geométrica
JRip	97,0	**81,7**	95,7	**89,0**
NNge	98,0	76,7	96,1	86,7
OneR	98,9	41,7	93,7	64,2
Prism	99,2	44,2	94,7	66,2
Ridor	95,6	68,3	93,1	80,8
ADTree	**99,2**	78,3	**97,3**	88,1
J48	97,7	55,5	93,9	73,6
RandomTree	98,0	63,3	94,9	78,8
REPTree	97,9	60,0	94,5	76,6
SimpleCart	98,0	65,0	95,1	79,8

Puede observarse en la Tabla IV que los porcentajes de exactitud obtenidos para la exactitud total y para los Aprobados son altos, no así para los que suspendieron y la media geométrica. Concretamente, los algoritmos que obtienen los valores máximos son: JRip (en el *ratio* de Suspendió/Abandonó y media geométrica), y ADTree (en el *ratio* de Suspendió/abandonó y exactitud).

En el segundo experimento, se han utilizado los ficheros con los mejores 15 atributos, que consiste en ejecutar nuevamente los 10 algoritmos de clasificación para poder comprobar cómo ha afectado la selección de atributos en la predicción. La Tabla V muestra los resultados de la validación cruzada (la media de las 10 ejecuciones) de los algoritmos de clasificación utilizando solamente los 15 mejores atributos.

Al comparar las Tablas IV y V se puede observar que los algoritmos han mejorado el porcentaje de exactitud al utilizar sólo los mejores atributos. Aunque hay algunos algoritmos que empeoran un poco, en general la tendencia es de mejora. De hecho, se obtienen unos valores máximos mejores a los obtenidos con todos los atributos. Nuevamente los algoritmos que obtienen estos valores máximos son el JRip (*ratio* Suspendió y media geométrica) y ADTree (*ratio* Aprobó y exactitud total).

A pesar de haber obtenido mejores resultados, todavía no se obtiene una buena clasificación de la clase minoritaria Suspendió/Abandonó, obteniendo como valor máximo sólo un 81,7% de acierto frente al 99,2% de acierto de la clase mayoritaria Aprobó. Esto puede ser debido a que los datos se encuentran muy desbalanceados. Esta característica de los datos es un hecho poco deseable y puede afectar negativamente en los resultados obtenidos al aplicar los algoritmos de clasificación, y es debido a que los algoritmos suelen centrarse en clasificar a los individuos de la clase mayoritaria para obtener un buen porcentaje de clasificación total y olvidar a los individuos de la clase minoritaria.

En el tercer experimento, se ha intentado solucionar o mitigar este problema del desbalanceo de los datos. Para ello, se han vuelto a utilizar los ficheros con los mejores 15 atributos, pero ahora los ficheros de entrenamiento han sido previamente re-balanceados con el algoritmo SMOTE. La tabla VI muestra los resultados de esta tercera prueba realizada.

Al analizar esta tabla y compararla con los anteriores resultados de las Tablas IV y V se observa que la mayoría de los algoritmos han aumentado su exactitud en predicción, obteniendo nuevos valores máximos en casi todas las medidas excepto en el porcentaje de exactitud total. En este caso, los algoritmos que han obtenido los mejores resultados han sido el algoritmo Prism, OneR y nuevamente el algoritmo ADTree.

Finalmente, otra forma distinta de abordar el problema de la clasificación de datos desbalanceados es realizar una clasificación sensible al costo. En la clasificación tradicional no se distingue si una de las clases a clasificar es más importante que otra, es decir, si una tiene un costo de clasificación diferente. Optimizar la tasa de clasificación sin tomar en cuenta el costo de los errores a menudo puede conducir a resultados no óptimos debido al alto costo que puede ocasionar la mala clasificación de una instancia minoritaria. De hecho, en nuestro problema en particular, estamos mucho más interesados en la clasificación de los alumnos de la clase Suspendió/Abandonó (clase minoritaria). Estos costos pueden ser incorporados al

algoritmo de forma que se puedan tener en cuenta durante la clasificación. Por ejemplo, si se tiene dos clases, los dos tipos de errores, Falsos Positivos y Falsos Negativos pueden tener diferente costo; o por el contrario, los dos tipos de clasificación correcta pueden tener diferentes beneficios. Los costos se pueden plasmar en una matriz de 2x2 en la cual los elementos de la diagonal principal representan los dos tipos de clasificación correcta y los elementos fuera de la diagonal representan los dos tipos de errores. La matriz de costos predeterminada (utilizada por todos los algoritmos de clasificación tradicional) en la que los costos de los errores son iguales sería: [0, 1; 1, 0], donde la diagonal principal (valores correctos) tienen ceros y en el resto (valores incorrectos) tiene unos, indicando que tienen igual costo o beneficio. En cambio si los costos de alguna de las clasificaciones erróneas o bien de alguno de los beneficios de las correctas son diferentes, se indicaría con un valor distinto a uno.

El software Weka permite realizar clasificación teniendo en cuenta el costo, para lo cual se utiliza el clasificador CostSensitiveClassifier y al cual se le asocia tanto la matriz de costo como el algoritmo clasificador a utilizar. Después de hacer varias pruebas con diferentes costos, se encontró que utilizando la matriz [0, 1; 4, 0], se obtuvieron los mejores resultados de clasificación, lo cual indica que al realizar la clasificación se tiene en cuenta que es 4 veces más importante clasificar de manera correcta los casos de Suspendió/Abandonó que los casos de Aprobó.

Finalmente, el cuarto y último experimento consistió en ejecutar los diez algoritmos de clasificación utilizando costos y los ficheros con los mejores 15 atributos. La Tabla VII muestra los resultados obtenidos.

Al comparar los resultados obtenidos en la Tabla VII con respecto a la Tabla VI, se observa que aunque ha empeorado un poco el porcentaje de aciertos de Aprobó y la exactitud total, por el contrario ha aumentado (obteniendo los valores máximos con respecto a todos los anteriores experimentos) tanto la media geométrica como el porcentaje de aciertos de Suspendió/Abandonó que es justamente lo que más nos interesa en este trabajo (detectar a los alumnos en riesgo). En este caso, los algoritmos que mejores resultados obtuvieron fueron Prism, JRip, ADTree y SimpleCart.

TABLA VI
VALIDACIÓN CRUZADA UTILIZANDO LOS 15 MEJORES ATRIBUTOS Y PREVIAMENTE BALANCEANDO LOS DATOS DE ENTRENAMIENTO CON EL ALGORITMO SMOTE

Algoritmo	%Aciertos Aprobó	%Aciertos Suspendió	%Exactitud Total	Media Geométrica
JRip	97,7	65,0	94,8	78,8
NNge	98,7	78,3	96,9	87,1
OneR	88,8	**88,3**	88,8	88,3
Prism	**99,8**	37,1	94,7	59,0
Ridor	97,9	70,0	95,4	81,4
ADTree	98,2	86,7	**97,2**	**92,1**
J48	96,7	75,0	94,8	84,8
RandomTree	96,1	68,3	93,6	79,6
REPTree	96,5	75,0	94,6	84,6
SimpleCart	96,4	76,7	94,6	85,5

TABLA VII
VALIDACIÓN CRUZADA UTILIZANDO LOS 15 MEJORES ATRIBUTOS Y CONSIDERANDO EL COSTO DE CLASIFICACIÓN

Algoritmo	%Aciertos Aprobó	%Aciertos Suspendió	%Exactitud Total	Media Geométrica
JRip	96,2	**93,3**	96,0	**94,6**
NNge	98,2	71,7	95,8	83,0
OneR	96,1	70,0	93,7	80,5
Prism	**99,5**	39,7	94,4	54,0
Ridor	96,9	58,3	93,4	74,0
ADTree	98,1	81,7	**96,6**	89,0
J48	95,7	80,0	94,3	87,1
RandomTree	96,6	68,3	94,0	80,4
REPTree	95,4	65,0	92,7	78,1
SimpleCart	97,2	90,5	**96,6**	93,6

VI. Interpretación de Resultados

En esta sección se van a mostrar y comentar algunos de los modelos de reglas o árboles de clasificación que han sido generados por los algoritmos que mejores resultados de clasificación han obtenido en la anterior etapa de experimentación: JRip (ver Tabla VIII), ADTree (ver Tabla IX), Prism (ver Tabla X) y SimpleCart (ver Tabla XI).

En las reglas de la Tabla VIII se observa en general que el algoritmo JRip descubre pocas reglas. Con respecto a los atributos que aparecen asociados a suspender son mayoritariamente referentes a notas, indicando que el alumno suspendió Física o que no se presentó a Humanidades, Matemáticas, Inglés o Taller de Lectura y Redacción. También aparecen otros atributos que indican que los alumnos que suspenden tienen una edad superior a 15 años o que pertenece a un determinado grupo (1M).

En el árbol de decisión de la Tabla IX se observa que sólo aparecen atributos de tipo nota con valores de no presentado, deficiente o regular. Se observa además que asignaturas como Humanidades y Ciencias Sociales, asignaturas relativamente sencillas de aprobar, aparecen en la parte alta del árbol.

En las reglas de la Tabla X se observa en general que el algoritmo Prism descubre una gran cantidad de reglas. Además, se observa que además de los atributos de tipo nota (con valores de no presentado o deficiente) aparece el atributo que indica que el alumno pertenece a un grupo en particular (1R, 1G, 1M o 1E).

El árbol de clasificación de la Tabla XI se observa en general que es más pequeño que el obtenido por el ADTree. También se observa que además de los atributos de tipo nota (con valores de no presentado, deficiente o regular) en Humanidades, Matemáticas, Inglés o Computación, también aparecen otros atributos como son si el nivel de motivación del alumno es bajo o no y a que semestre y grupo en particular pertenece el alumno (1R, 1G, 1M).

Finalmente, es importante reseñar que no se ha detectado un consenso entre los anteriores algoritmos de clasificación sobre la existencia de un único factor que más influya en el fracaso de los estudiantes. En cambio, si se pueden considerar el siguiente grupo de factores (que son los que más aparecen en los modelos obtenidos) como los más influyentes: Deficiente o No Presentado en Física 1, Deficiente o No Presentado en Matemáticas, No Presentado en Humanidades 1, Deficiente en Inglés 1, No Presentado en Taller de Lectura y Redacción, Deficiente en Ciencias Sociales, Edad de más de 15 años y Nivel de motivación regular.

TABLA VIII
REGLAS OBTENIDAS USANDO JRip, LOS 15 MEJORES ATRIBUTOS Y CONSIDERANDO EL COSTO DE LA CLASIFICACIÓN

(Física 1 = Deficiente) => Estado Académico = Suspendió
(Humanidades 1 = No Presentado) => Estado Académico = Suspendió
(Matemáticas 1 = No Presentado) => Estado Académico = Suspendió
(Inglés 1 = Deficiente) and (Física 1 = No Presentado) => Estado Académico = Suspendió
(Taller de Lectura y Redacción 1 = No Presentado) => Estado Académico = Suspendió
(Ciencias Sociales 1 = Deficiente) and (Edad = más de 15) and (Semestre y Grupo = 1M) => Estado Académico = Suspendió
=> Estado Académico = Aprobó

TABLA IX
ÁRBOL OBTENIDO USANDO ADTree, LOS 15 MEJORES ATRIBUTOS Y CONSIDERANDO EL COSTO DE LA CLASIFICACIÓN

```
: -0,465
| (1) Humanidades 1 = No Presentado: 1,824
| (1) Humanidades 1!= No Presentado: -0,412
|   | (2) Física 1 = Deficiente: 1,415
|   | (2) Física 1 != Deficiente: -0,632
|   |   | (9) Taller de Lectura y Redacción 1 = No Presentado: 1,224
|   |   | (9) Taller de Lectura y Redacción 1 != No Presentado: -0,52
| (3) Ciencias Sociales 1 = Deficiente: 1,689
| (3) Ciencias Sociales 1 != Deficiente: -0,245
|   | (4) Inglés 1 = Deficiente: 1,278
|   | (4) Inglés 1 != Deficiente: -0,322
|   |   | (5) Matemáticas 1 = No Presentado: 1,713
|   |   | (5) Matemáticas 1 != No Presentado: -0,674
|   |   | (6) Ciencias Sociales 1 = No Presentado: 1,418
|   |   | (6) Ciencias Sociales 1 != No Presentado: -0,283
|   |   |   | (8) Inglés 1 = No Presentado: 1,313
|   |   |   | (8) Inglés 1 != No Presentado: -0,695
| (7) Matemáticas 1 = Deficiente: 0,758
|   | (10) Humanidades 1 = Regular: -0,473
|   | (10) Humanidades 1 != Regular: 0,757
| (7) Matemáticas 1 != Deficiente: -0,315
Legend: -ve = Aprobó, +ve = Suspendió
```

TABLA X
REGLAS (DE TIPO SUSPENDIÓ) OBTENIDAS USANDO Prism, LOS 15 MEJORES ATRIBUTOS Y CONSIDERANDO EL COSTO DE LA CLASIFICACIÓN

If Taller de Lectura y Redacción 1 = No Presentado then Suspendió
If Ciencias Sociales 1 = No Presentado and Humanidades 1 = No Presentado then Suspendió
If Humanidades 1 = No Presentado and Matemáticas 1 = Deficiente then Suspendió
If Matemáticas 1 = No Presentado and Calificación de EXANI I = Muy Deficiente then Suspendió
If Taller de Lectura y Redacción 1 = Deficiente and Ciencias Sociales 1 = Deficiente then Suspendió
If Matemáticas 1 = No Presentado and Inglés 1 = No Presentado then Suspendió
If Inglés 1 = Deficiente and Computación 1 = Regular then Suspendió
If Ciencias Sociales 1 = Deficiente and Física 1 = Deficiente then Suspendió
If Computación 1 = Deficiente and Matemáticas 1 = Deficiente then Suspendió
If Ciencias Sociales 1 = No Presentado and Semestre y Grupo = 1º R then Suspendió
If Inglés 1 = Deficiente and Semestre y Grupo = 1º G then Suspendió
If Humanidades 1 = Deficiente and Matemáticas 1 = Deficiente and Semestre y Grupo = 1º M then Suspendió
If Humanidades 1 = No Presentado and Semestre y Grupo = 1º E then Suspendió

VII. CONCLUSIONES.

En este trabajo, se realizaron un conjunto de experimentos con el objetivo de conseguir predecir con un buen grado de exactitud el estado académico de los estudiantes al final del primer semestre mediante la utilización de algoritmos de clasificación.

TABLA XI
ÁRBOL OBTENIDO USANDO SimpleCart, LOS 15 MEJORES ATRIBUTOS Y CONSIDERANDO EL COSTO DE LA CLASIFICACIÓN

```
Humanidades 1 = (No Presento)|(Deficiente)
| Matemáticas 1 = (No Presentado)|(Deficiente): Suspendió
| | Semestre y Grupo = (1º M): Suspendió
Humanidades 1! = (No Presentado)|(Deficiente)
| Inglés 1 = (Deficiente)|(No Presentado)
| | Semestre y Grupo = (1º G): Suspendió
| | | Nivel de Motivación = (Regular): Suspendió
| | | Nivel de Motivación! = (Regular): Aprobó
| Inglés 1! = (Deficiente)|(No Presentado)
| | Ciencias Sociales 1 = (Deficiente)
| | | Semestre y Grupo = (1º R): Suspendió
| | | Semestre y Grupo! = (1º R): Aprobó
| | Ciencias Sociales 1! = (Deficiente): Aprobó
```

Como se ha podido ver, este objetivo no es sencillo conseguirlo, debido a que no sólo se trata de un conjunto de datos desbalanceado, sino que es un problema multifactorial.

Es importante, comentar que una tarea muy importante en este trabajo, fue la recopilación de la información y el pre-procesado de los datos, ya que la calidad y fiabilidad de la información afecta de manera directa en los resultados obtenidos. Es una tarea ardua, que implica invertir mucho tiempo y disposición de quien esté a cargo de realizarla. En concreto se tuvo que realizar la captura de los datos de la encuesta aplicada, además de hacer la integración de datos de tres fuentes diferentes para formar el conjunto de datos final.

Respecto a los resultados de clasificación de las diferentes pruebas, las principales conclusiones que hemos obtenido son:

- Se ha mostrado que los algoritmos de clasificación pueden utilizarse con éxito para predecir el rendimiento académico de los estudiantes.
- Se ha mostrado la utilidad de las técnicas de selección de características cuando se dispone de muchos atributos, consiguiendo mejorar la clasificación de los algoritmos al utilizar un conjunto reducido de 15 atributos de entre los 77 disponibles inicialmente.
- Se ha mostrado dos formas distintas de abordar el problema de clasificación de datos desbalanceados, tanto re-balanceando los datos como considerando distintos costos de clasificación y aplicando una matriz de costos. Ambas formas han conseguido mejorar la clasificación, aunque para nuestro problema en particular la matriz de costos ha obtenido los mejores resultados de clasificación de la clase minoritaria que es lo que más nos interesa.

Respecto del conocimiento extraído de los modelos de clasificación obtenidos, las principales conclusiones son:

- La utilización de algoritmos de clasificación de tipo "caja-blanca" permiten obtener modelos comprensibles por un usuario no experto en minería de datos en procesos de toma de decisiones. En nuestro caso el objetivo final es poder detectar los alumnos con problemas o tendencia a Suspender/Abandonar para intentar impedirlo a tiempo.
- Con respecto a los factores que más han aparecido en los modelos obtenidos: Las notas de las asignaturas del semestre son las que aparecen en mayor medida en las salidas de los algoritmos de clasificación,

siendo las más importantes los que obtuvieron unas notas deficientes o no se presentaron a las asignaturas de Física 1, Humanidades 1, Matemáticas 1 e Inglés 1. Además otros atributos que han aparecido en los modelos han sido la edad (particularmente los mayores de 15), tener hermanos (particularmente 1), el grupo al que asistió, el nivel (regular) de motivación por estudiar, el no presentarse al Taller de Lectura y Redacción, el vivir en una ciudad grande (particularmente en una comunidad de más de 20 mil habitantes) y el considerar que la asignatura más difícil sea Matemáticas. Llama la atención que la nota deficiente de una asignatura como Humanidades, que generalmente es aprobada por la mayoría de los estudiantes aparezca en los modelos obtenidos como un factor relacionado con el fracaso de los estudiantes. Hay que indicar también que en este estudio hemos utilizado las notas anteriores de los estudiantes y no nos hemos centrado sólo en los atributos sociales, culturales y demográficos debido a dos motivos. Primero, los resultados de clasificación que obteníamos si eliminábamos los atributos de notas anteriores empeoraban muchísimo. Segundo, las notas anteriormente obtenidas por los estudiantes para predecir el fracaso es un recurso muy utilizado en trabajos similares [21], [22].

A partir de los modelos de reglas y los árboles de decisión generados por los algoritmos de minería de datos se puede implementar un sistema que alerte al profesor sobre los estudiantes que potencialmente se encuentren en riesgo de suspender o abandonar, así como a sus padres. Como ejemplo de posibles acciones que pueden servir de ayuda a los estudiantes en riesgo, se propone, que una vez que se detecte un alumno con riesgo, se les asigne un profesor-tutor para que les brinde un apoyo tanto académico como motivador y orientador para intentar evitar el fracaso del alumno.

Finalmente, como siguiente paso en nuestra investigación vamos a realizar más pruebas utilizando otros datos que estamos actualmente capturando y pre-procesando. Y como líneas de trabajo futuro, podemos destacar las siguientes:

- Desarrollar un algoritmo propio de clasificación/predicción basado en programación genética basada en gramática para poder compararlo con los resultados de algoritmos clásicos y obtener mejores resultados de predicción.
- Intentar predecir el fracaso o abandono de los alumnos lo antes posible, mientras más temprano mejor, para poder detectar a los alumnos en peligro a tiempo antes de que ya sea tarde.
- Proponer métodos para ayudar a los alumnos detectados dentro del grupo de riesgo de suspender o abandonar. Posteriormente comprobar que porcentaje de las veces fue posible evitar que un alumno detectado a tiempo fracasara o abandonara.

AGRADECIMIENTOS

Los autores agradecen el soporte económico proporcionado por el ministerio de educación, ciencia e innovación (Proyecto TIN-2011-22408) y la Junta de Andalucía (Proyecto P08-TIC-3720).

REFERENCIAS

[1] L. A. Alvares Aldaco, “Comportamiento de la Deserción y Reprobación en el Colegio de Bachilleres del Estado de Baja California: Caso Plantel Ensenada”, X Congreso Nacional de Investigación Educativa. México, 2009.

[2] F. Araque, C. Roldán, A. Salguero, “Factors Influencing University Drop Out Rates”, Computers & Education, vol. 53, pp. 563–574, 2009.

[3] M. N. Quadril and N. V. Kalyankar, “Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques”, Global Journal of Computer Science and Technology, vol. 10, pp. 2-5, 2010.

[4] C. Romero and S. Ventura, “Educational data mining: A Survey From 1995 to 2005”, Expert System with Applications, vol. 33, pp. 135-146, 2007.

[5] C. Romero and S. Ventura, “Educational Data mining: A Review of the State of the Art”, IEEE Transactions on Systems, Man, and Cybernetics, 2010.

[6] S. Kotsiantis, K. Patriarcheas and M. Xenos, “A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´ Performance in Distance Education”, Knowledge Based System, vol. 23, no. 6, pp. 529-535, 2010.

[7] J. Más-Estellés, R. Alcover-Arándiga, A. Dapena-Janeiro, A. Valderruten-Vidal, R. Satorre-Cuerda, F. Llopis-Pascual, T. Rojo-Guillén, R. Mayo-Gual, M. Bermejo-Llopis, J. Gutiérrez-Serrano, J. García-Almiñana, E. Tovar-Caro, E. Menasalvas-Ruiz, “Rendimiento Académico de los Estudios de Informática en Algunos Centros Españoles”, XV Jornadas de Enseñanza Universitaria de la Informática, Barcelona, Reporte de Conferencia, 2009.

[8] S. Kotsiantis, “Educational Data Mining: A Case Study for Predicting Dropout – Prone Students”, Int. J. Knowledge Engineering and Soft Data Paradigms, vol. 1, no. 2, pp. 101–111, 2009.

[9] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, “Dropout Prediction in e-learning Courses through the Combination of Machine Learning Techniques”, Computers & Education, vol. 53, pp. 950–965, 2009.

[10] A. Parker, “A Study of Variables That Predict Dropout From Distance Education”, International Journal of Educational Technology, vol. 1, no. 2, pp. 1-11, 1999.

[11] Aluja, T., "La minería de datos, entre la estadística y la inteligencia artificial", Quaderns d'Estadística i Investigació Operativa, vol. 25, num 3., pp. 479-498, 2001.

[12] M. M. Hernández, “Causas del Fracaso Escolar”, XIII Congreso de la Sociedad Española de Medicina del Adolescente, pp.1-5. 2002.

[13] E. Espíndola, A. León, “La Deserción Escolar en América Latina un Tema Prioritario Para la Agenda Regional”, Revista Iberoamericana de Educación, no. 30, pp. 1-17, 2002.

[14] I. H. Witten and F. Eibe, “Data Mining, practical Machine Learning Tools and Techniques”, Second Edition, Morgan Kaufman Publishers, 2005.

[15] M. A. Hall and G. Holmes, “Benchmarking Attribute Selection Techniques for Data Mining”, Technical Report 00/10, University of Waikato, Department of Computer Science, Hamilton, New Zealand, Julio 2002. Available: http://www.cs.waikato.ac.nz/~ml/publications/2000/00MH-GH-Benchmarking.pdf.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, W.P. Kegelmeyer, “Synthetic Minority Over-sampling Technique”, Journal of Artificial Intelligence Research, 2002, 16:321-357.

[17] J. Cendrowska, “PRISM: An algorithm for inducing modular rules”, International Journal of Man-Machine Studies, vol. 27, no. 4, pp. 349-370, 1987.

[18] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufman Publishers, 1993.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, “Classification and Regression Trees”, Chapman & Hall, New York, 1984.

[20] Y. Freund and L. Mason, “The Alternating Decision Tree Algorithm”, Proceedings of the 16th International Conference on Machine Learning, pp. 124-133, 1999.

[21] L. Fourtin , D. Marcotte, P. Potvin, E. Roger, J. Joly, "Typology of students at risk of dropping out of school: Description by personal, family and school factors", European Journal of psychology of education, vol. XXI, no. 4, pp. 363-383, 2006.
[22] L. G. Moseley, D. M. Mead, "Predicting who will drop out of nursing courses: A machine learning exercise", Nurse Education Today, vol. 28, 469-475, 2008.

Carlos Márquez Vera es profesor de la Universidad Autónoma de Zacatecas, México y doctorando en la Universidad de Córdoba, España. Su área de interés principal es la minería de datos educativa.

Cristóbal Romero Morales es profesor titular del departamento de Informática de la Universidad de Córdoba en España. Es doctor en Informática por la Universidad de Granada desde el año 2003. Su área de interés principal es la aplicación de minería de datos educativa.

Sebastián Ventura Soto es profesor titular del departamento de Informática de la Universidad de Córdoba en España. Es doctor en Ciencias por la Universidad de Córdoba desde el año 1996. Su área de interés principal es soft-computing y sus aplicaciones.