Reglas que describen la deserción y permanencia en los estudiantes de la UAP Tianguistenco de la UAEM

Guillermo García Lambert*, René Arnulfo García-Hernández* y Yulia Ledeneva*

Recepción: 6 de junio de 2013
Aceptación: 13 de diciembre de 2013

***Unidad Académica Profesional Tianguistenco, Universidad Autónoma del Estado de México, México. Correos electrónicos: ggarcial@uaemex.mx; renearnulfo@hotmail.com y yledeneva@yahoo.com**
Se agradecen los comentarios de los árbitros de la revista.

Resumen. Se pretende encontrar cuál es el conjunto de reglas de conocimiento que pueden extraerse de aquellos estudiantes que han desertado o que permanecen en sus estudios universitarios tres años después de su ingreso. Se utilizó una base de datos inicial con 206 factores y 305 estudiantes de cuatro licenciaturas de la UAP Tianguistenco de la UAEM. Mediante árboles de decisión, fue posible determinar que con sólo 12 factores en 19 reglas se puede saber, con un 82.6% de soporte, si un estudiante tiene riesgo de desertar o no de sus estudios en los tres años posteriores.

Palabras clave: minería de datos en estudiantes, árbol de decisión C4.5, eficiencia terminal de estudiantes universitarios, deserción de estudiantes universitarios.

Rules that describe the dropout and permanence of UAP Tianguistenco (UAEM) student

Abstract. In this research, we are focused on finding what the set of knowledge rules are, that we can get from students who have dropped out or continued in college three years later after admission. A preliminary database was used with 206 factors and 305 students of four different degrees from the UAP Tianguistenco of the UAEM. It was possible to determine by applying decision trees, that with only 12 factors in 19 rules, it is possible to know, with 82.6% support, if a student is at risk of dropping out or not from his studies in the next three years.

Key words: data mining of students, C4.5 decision tree, graduation rate of college students, college dropouts.

Introducción

El bajo índice de eficiencia terminal (cantidad de estudiantes titulados) es un problema nacional dada la gran cantidad de recursos económicos y humanos que el gobierno invierte para ello. La deserción en los estudios universitarios forma parte de este problema, que según la ANUIES (Asociación Nacional de Universidades e Instituciones de Educación Superior) en 2000 fue de 39% en México (ANUIES, 2005). En el caso de la Universidad Autónoma del Estado de México (UAEM) se reportó una eficiencia terminal de técnico superior y licenciatura por programa educativo 2004-2005/2008-2009 por cohorte de 53.7%, es decir, de 7 438 estudiantes que ingresaron a nivel superior, 3 995 egresaron en los tiempos establecidos por el programa educativo. Ahora bien, si se considera el egreso con las generaciones rezagadas entonces se tiene una eficiencia terminal global[1]

1. La eficiencia terminal global incluye estudiantes egresados de otras generaciones por lo que en algunos casos es mayor a 100. La eficiencia terminal por cohorte se calcula con el nuevo ingreso de acuerdo con la duración del programa educativo (UAEM, 2011).

de 68.4% (UAEM, 2011). De acuerdo con el estudio de Wietse de Vries *et al.* (2011) los datos reportados por la ANUIES sobre la eficiencia terminal han llegado a bajar 20% para una eficiencia por cohorte en algunos años.

Los estudiantes de nivel superior no sólo dependen de los conocimientos académicos, sino que además existen factores socio-económicos y familiares que influyen en su formación académica (Pontón, 2006). En específico, E. Tomul (2009) menciona que las variables familiares, la educación de los padres y el logro académico en el área de matemáticas parecen afectar el logro estudiantil.

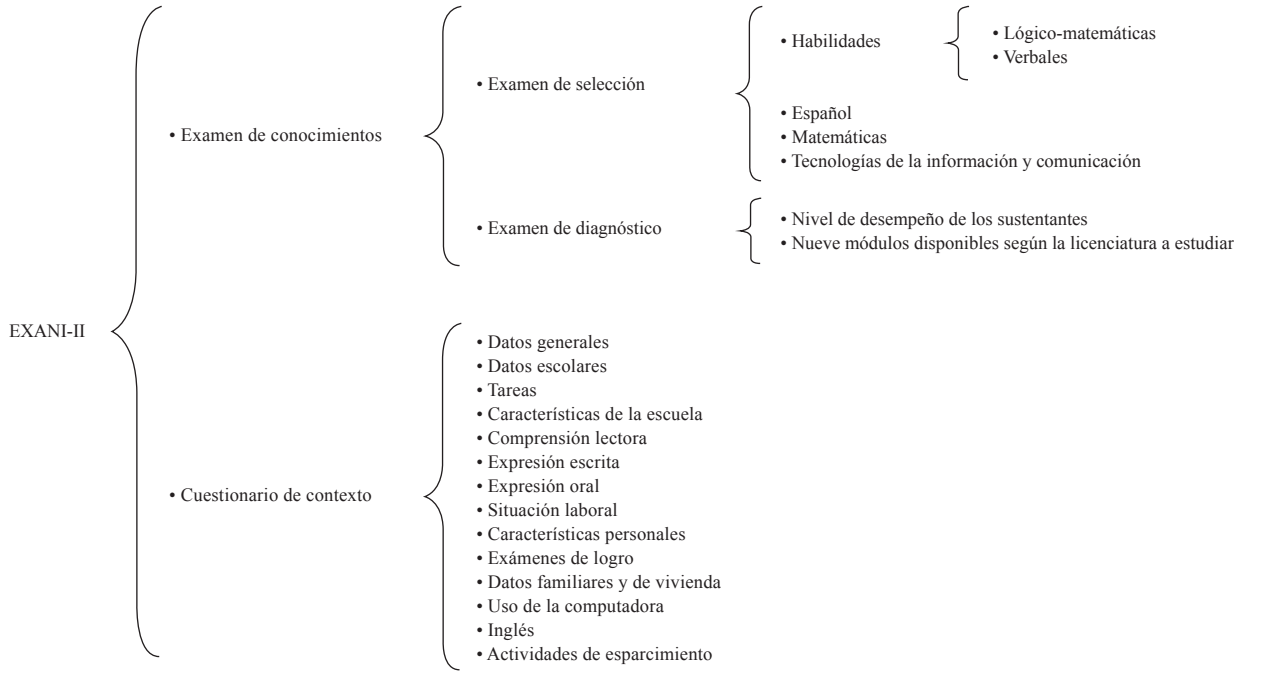
Por su parte, J. Murillo (2003) provee una panorámica de una investigación iberoamericana sobre eficacia escolar, donde se planteó encontrar la relación entre determinados factores escolares y el rendimiento de los estudiantes, de los cuales resaltan el ambiente escolar, la infraestructura, los recursos de la escuela y el ambiente del aula.

Además, existen relaciones entre la motivación del logro, actitud y rendimiento académico en los estudiantes, según lo afirma el estudio que realizó K. Bakar (Kamariah *et al.,* 2010) en Malasia. En dicho estudio determinan algunos factores significantes que influyen en el logro estudiantil como la motivación, la actitud hacia el aprendizaje, el género y el grupo étnico al que pertenece el estudiante.

Como puede apreciarse, esta problemática se ha estudiado en diversos países y niveles educativos, también es posible ver que es multifactorial, puesto que son diversos los factores que afectan al estudiante. No obstante, no todos los factores censados son determinantes para este estudio, por lo que es indispensable censar inicialmente el mayor número de factores posibles con el objetivo de encontrar cuáles son importantes.

Se propone utilizar la información recabada por el examen de admisión EXANI-II (Examen Nacional de Ingreso a la Educación Superior) que aplica el CENEVAL (Centro Nacional de Evaluación para la Educación Superior) a los estudiantes de bachillerato para su ingreso a los estudios superiores de la UAEM. El EXANI-II está compuesto por un examen de conocimientos y un cuestionario de contexto. El examen de conocimientos se compone a su vez de uno de selección y uno de diagnóstico. El examen de selección evalúa los conocimientos del sustentante en áreas de español, matemáticas, tecnologías de la información y comunicación, así como habilidades tanto lógico-matemáticas como verbales. El examen de diagnóstico se aplica de acuerdo con la licenciatura a la que desea ingresar. A diferencia del examen de conocimientos, el cuestionario de contexto soló recaba información socioeconómica, datos generales, datos escolares, situación laboral, características personales, datos familiares, de vivienda, entre otros. La figura 1 muestra las secciones principales del examen EXANI-II de 2008. En específico, el cuestionario de contexto está conformado

Figura 1. **Estructura del contenido del examen EXANI-II.**



Fuente: elaboración propia.

por 206 preguntas distribuidas en 14 secciones, las cuales censan información acerca del aspirante.

Una vez que el estudiante ha sido aceptado, el departamento de control escolar registra su historial académico cada semestre a lo largo de sus estudios, en el cual se señala a quienes se dieron de baja voluntariamente o por no cumplir con los requisitos para permanecer; ambos casos son considerados en esta investigación como deserción escolar.

Por medio de la gran cantidad de factores recabados por el cuestionario de contexto del EXANI-II y la información que proporciona control escolar de los estudiantes que desertaron, se busca determinar cuáles de los factores y en qué porcentaje son característicos de los estudiantes que desertan y de aquellos que permanecen. Además, se pretende determinar la jerarquía que cada factor guarda dentro del subconjunto encontrado. Cabe comentar que este problema puede parecer sencillo a primera vista; sin embargo, para escoger el mejor subconjunto se tendría que revisar $\sum_{n=1}^{206} n! = 5.627257e + 388$ subconjuntos, lo cual es impráctico con los sistemas de cómputo actuales, por lo que se hace imprescindible utilizar técnicas estadísticas o de minería de datos para analizar el subconjunto de datos.

Una de las técnicas estadísticas que se ha empleado para este estudio es la *regresión lineal* (Pontón, 2006; Tomul, 2009). Si bien permite modelar la relación que existe entre una variable dependiente y las variables independientes, no es posible para las variables descritas con valores categóricos. Por ejemplo, para la pregunta ¿Con qué frecuencia se impone fechas límite para terminar trabajos importantes? Las respuestas pueden ser "siempre", "frecuentemente", "algunas veces" y "nunca". A pesar de esto, creemos que las variables categóricas contienen información relevante para nuestro problema.

Por otro lado, en la regresión lineal se asume que la variable dependiente es causada por todas las independientes y, que por serlo, no tienen ninguna relación entre sí. En cambio, en esta investigación estamos interesados en descartar primero aquellas variables que no presenten una alta incidencia respecto a nuestro objeto de estudio.

La hipótesis es entonces que las bases de datos de estudiantes, obtenidas del cuestionario de contexto EXANI-II y de control escolar proveen información suficiente para extraer subconjuntos de factores que permitan caracterizar a los estudiantes que desertan de aquellos que no lo hicieron, lo cual es posible hacer en tiempos razonables mediante las técnicas de minería de datos. En específico, se plantea utilizar los árboles de decisión debido a que permite analizar variables numéricas, nominales y categóricas, incluso, que haya faltantes de información como sucede en este caso cuando el estudiante no contestó una pregunta.

1. Estado del arte

De manera resumida, se muestran algunos estudios que se han realizado sobre esta problemática. Se presenta el modelo de análisis que emplearon en sus estudios y los factores que, de acuerdo con sus conclusiones, son determinantes o influyen en los estudiantes (ver anexo, tabla A1).

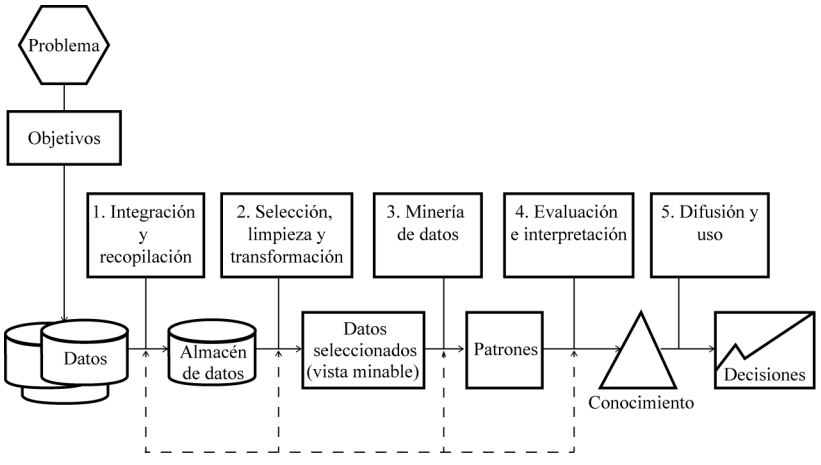
2. Propuesta metodológica

Se utilizó el descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés) como la metodología para llevar a cabo los experimentos de manera ordenada y sistemática (Fayyad *et al.,* 1997). El KDD aplica iterativamente las fases de *a*) integración y recopilación, *b*) selección, limpieza y transformación, *c*) minería de datos, *d*) evaluación e interpretación y *e*) difusión y uso, como se ve en la figura 2 (Hernández Orallo *et al.,* 2004).

2. 1. Fase de integración y recopilación

Se determinan las fuentes de información para formar posteriormente la base de datos del objeto de estudio. Para esto se utilizan los datos recabados de los estudiantes que presentaron el examen EXANI-II en 2008 para ingresar a las

Figura 2. **Diagrama de descubrimiento de conocimiento en bases de datos.**



Fuente: Hernández Orallo *et al.,* 2004.

carreras de Ingeniero en *Software* (IS), Ingeniero en Producción Industrial (IPI), Ingeniero en Plásticos (IP) y la Licenciatura en Seguridad Ciudadana (LCS) de la Unidad Académica Profesional Tianguistenco de la UAEM. En segunda instancia, se encuentra la información registrada por el departamento de control escolar a lo largo de 3 años. Cabe señalar que al momento de realizar los experimentos de la investigación, los estudiantes que ingresaron en 2008 se encontraban en el sexto semestre de diez que contemplan las cuatro carreras.

2. 2. Fase de selección, limpieza y transformación

Se busca obtener la primera vista minable de la base de datos adecuada, por lo que es necesario seleccionarlos y luego transformarlos en representaciones adecuadas para la minería de datos. Una vez realizada la unión del EXANI-II y control escolar se obtuvo una base de datos inicial con 305 estudiantes y 214 factores (206 del EXANI-II y 8 de control escolar). La base original de control escolar está compuesta por los siguientes atributos:

a) promedio general
b) promedio del periodo cursado
c) número de materias acreditadas
d) créditos obtenidos
e) materias reprobadas
f) promedio anual
g) si el estudiante tuvo beca
h) si el estudiante ha sido dado de baja, ya sea por el sistema o baja voluntariamente (atributo utilizado para formar las dos clasificaciones).

Empleando el atributo de baja se dividió la base de datos en dos grupos: quienes fueron dados de baja y quienes continúan con sus estudios, por lo que realmente se tienen 7 factores de control escolar y 213 factores en total.

Se eliminaron aquellos atributos que no son relevantes en este caso. Por ejemplo, el nombre del estudiante, ya que no se considera un factor en su desempeño; de no hacerlo es posible obtener como resultado que el apellido García tiene una alta probabilidad de ser dado de baja, que se debe a que es frecuente entre los mexicanos. En total se eliminaron los siguientes 20 atributos del EXANI-II: folio, clave eventual del estudiante, tipo de examen, tipo de registro, aplicador, fecha de aplicación, clave institucional, identificación, nombre, fecha de nacimiento (pero no la edad), CURP (Clave Única de Registro de Población), domicilio, teléfono, identificación del bachillerato (realmente está compuesta por tres atributos: nombre de la institución, clave, registro), clave de la escuela al ingresar, clave del periodo de ingreso, nombre de la escuela, carrera deseada (los estudiantes ya se tienen separados por las cuatro carreras).

En el caso de número de créditos aprobados (valoración curricular que se le da a cada materia) se normalizó con el número de créditos totales de cada carrera con el objetivo de procesar las cuatro carreras en conjunto. De esta manera se obtuvo la base de datos con las características mostradas en la tabla 1.

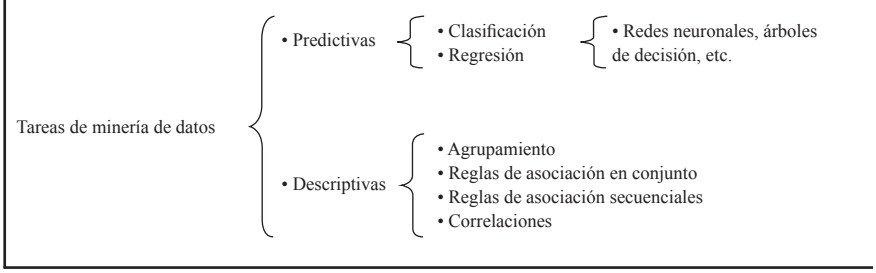
2. 3. Fase de minería de datos

Se pretende conseguir patrones y por ende producir conocimiento nuevo que pueda utilizar el usuario (Hernández Orallo *et al.,* 2004). Las tareas de la minería de datos se clasifican en predictivas y descriptivas. Las predictivas se subdividen en tareas de clasificación y regresión, donde las primeras pueden utilizar algoritmos como redes neuronales o árboles de decisión (figura 3). También existen diferentes paradigmas detrás de las técnicas utilizadas para esta fase: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y varios tipos de métodos basados en núcleos (Hernández Orallo *et al.,* 2004).

Tabla 1. **Bases de datos utilizadas para la experimentación.**

Base de datos	Factores (atributos)	Estudiantes (instancias)	
EXANI-II	186	62 dados de baja	243 permanecen estudiando
EXANI-II + Control	186 + 7		

Figura 3. **Diagrama parcial de las tareas y sub-tareas de la minería de datos.**



2. 3. 1. Árboles de decisión

Los árboles de decisión son modelos de predicción y parte de las técnicas de aprendizaje para la representación del conocimiento. Están basados en el principio de "divide y vencerás", los cuales Morgan y Sonquist desarrollaron en 1963.

Con ellos se obtienen modelos comprensibles de decisión a partir de una muestra de datos disponible que se representan de manera simbólica, o bien, en forma de reglas escritas de la forma si *antecedente*(*s*), entonces *consecuente*. El algoritmo C4.5 forma parte de la familia de los árboles de decisión que realizan su análisis de arriba hacia abajo (López, 2011), los cuales evalúan la información en cada caso por medio de los criterios de entropía, ganancia o proporción de ganancia, según sea el caso.

El problema de construir un árbol de decisión puede expresarse de manera recursiva. Primero, se selecciona un atributo para colocarlo como nodo raíz y hacer una ramificación para cada posible valor. Esto divide el conjunto de ejemplo en subconjuntos, uno por cada valor del atributo (Witten y Frank, 2005). En específico, el algoritmo C4.5 trabaja de la siguiente manera (Wu *et al.,* 2008):

Dado un conjunto de entrenamiento *S*, se genera un árbol inicial usando el algoritmo divide y vencerás como sigue:

a) si el número de casos en *S* pertenece a la misma clase o *S* es pequeña, el árbol sería una hoja etiquetada con la clase más frecuente en *S*.

b) de lo contrario, se elige el atributo con mayor ganancia de información. Este análisis se hace desde la raíz del árbol con una rama para cada respuesta del atributo.

c) se divide *S* en subconjuntos correspondientes S_1, S_2,…, S_n de acuerdo con el resultado de cada caso y se aplica el mismo procedimiento de forma recursiva para cada subconjunto S_i.

2. 4. Evaluación e interpretación

Los patrones hallados dentro de esta fase deben cumplir con tres criterios: ser precisos, inteligibles e interesantes. Es decir, deben ser comprendidos, tener utilidad y deben ser novedosos. Las reglas formadas de manera si *antecedente*, entonces *consecuente* cumplen con el principio de ser comprendidos por el humano. La precisión viene dada por el número de casos que cubre la regla; entonces, la novedad será por el conjunto de patrones no vistos antes, que a su vez tendrán varias aplicaciones.

Para saber qué tan significativos son los resultados reportados se utiliza la validación cruzada con *n*-iteraciones, donde se subdivide aleatoriamente los alumnos de la base de datos en *n* subconjuntos para los cuales uno de ellos se emplea para probar qué tan bien aprende a clasificar y el resto (de manera conjunta) para entrenar el aprendizaje del clasificador y se repite este proceso *n* veces hasta obtener así *n* índices de precisión independientes entre sí. El resultado final es el promedio de las *n* evaluaciones, que normalmente es $n = 10$.

2. 5. Difusión y uso

Una vez construido y validado el conocimiento, como lo menciona Hernández Orallo *et al.* (2004) puede usarse principalmente con dos finalidades: *a*) para que un analista recomiende acciones basándose en el modelo y *b*) en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos.

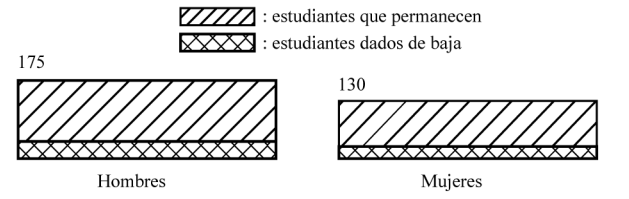
3. Experimentación

En el primer experimento sólo se consideró la información de la base de datos correspondiente al examen EXANI-II con el objetivo de ver cuáles de los factores recabados inicialmente (3 años antes de que pueda o no ser dado de baja) inciden en su permanencia o baja de sus estudios. En el segundo se considera a toda la base de datos y así saber si la información obtenida por control escolar sobre el desempeño dentro de la universidad permite obtener reglas con un mejor soporte.

Antes de realizar los experimentos, se exponen algunos datos estadísticos obtenidos de la base de datos que se va a trabajar. La distribución de la clase por sexo es de 57.37% hombres y de 42.63% mujeres. Sin embargo, como se puede ver en la gráfica 1, esta condición no contribuye como factor para saber si será dado de baja o no.

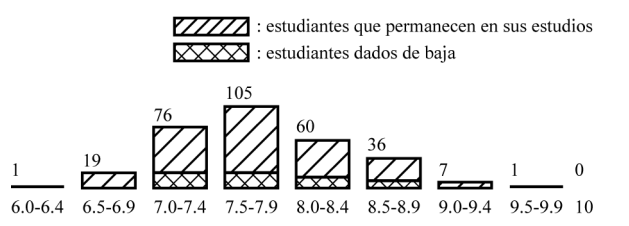
En la gráfica 2 se observa que aunque aumente el promedio general del bachillerato, siguen existiendo estudiantes dados de baja. Por ejemplo, la proporción de estudiantes que

Gráfica 1. **Distribución por sexo de los estudiantes dados de baja y que permanecen en sus estudios.**



Fuente: elaboración propia.

Gráfica 2. **Promedio general de bachillerato.**



Fuente: elaboración propia.

tenían un promedio entre 7.5 y 7.9, y que han sido dados de baja, es mucho menor que la de aquellos que tiene promedio mayor a 8.0. Es decir, si un estudiante tiene un promedio entre 7.5 y 7.9, tiene menos probabilidad de ser dado de baja que aquellos que tienen uno entre 8.0 y 8.4 o entre 8.5 y 8.9

3. 1. Primer experimento

De acuerdo con la metodología de KDD y la base de datos del EXANI-II se generó el árbol de decisión con el algoritmo C4.5 utilizando el *software* de minería de datos Weka (Hall *et al.,* s. f.), el cual se muestra en el anexo, figura A1 donde se encuentra la caracterización de los estudiantes tanto de aquellos que desertaron como de aquellos que permanecen en sus estudios, y que cada camino que puede recorrerse a partir del nodo raíz a cada uno de los nodos hoja forman, en esa jerarquía, una regla que permite saber cuáles son los factores determinantes para que un subgrupo de estudiantes pertenezca a una clase. Los nodos (hoja en color negro) corresponden a las reglas de los estudiantes dados de baja y los nodos hoja en color blanco de los estudiantes que permanecen en sus estudios. Para saber qué tan precisa es cada regla en particular (patrón) se incluye al final de cada nodo hoja, entre paréntesis, el soporte de la regla en porcentaje con el objetivo de cuantificarla independientemente del número de estudiantes. Según el árbol formado, es posible caracterizar a los estudiantes que fueron dados de baja en 11 reglas y los que no lo fueron en 12 reglas.

Cabe señalar que la suma de los soportes de las reglas de una clase no necesariamente tiene que dar 100%, debido a que el algoritmo reporta a quienes pudieron clasificarse correctamente y no el total de los estudiantes en esa clase. Por ejemplo, de los 62 estudiantes dados de baja, el algoritmo clasificó correctamente en esta clase a 41 (66% de los casos). Entonces, el soporte de la regla se calculó refiriéndose a los estudiantes dados de baja (62) y no a los correctamente clasificados (41).

La regla más importante encontrada para los estudiantes que tienen riesgo de ser dados de baja tiene un soporte de 12.9% de los casos, la cual se describe a continuación:

a) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, no sabe escribir apuntes en inglés cuando hablan varios expositores, frecuentemente pone fechas límite para terminar trabajos importantes, no vive con su pareja, el ingreso mensual del padre o tutor es de entre $5 001 a $10 000 y no es capaz de hacer anuncios públicos en inglés, entonces sí será dado de baja.

Las siguientes reglas son para los estudiantes que tiene un riesgo de ser dados de baja con un soporte de 11.3%, 11.3% y 9.7% respectivamente:

b) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene más de una computadora en casa, y es hábil para escribir y mandar correos electrónicos, entonces sí será dado de baja.

c) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, no sabe escribir apuntes en inglés cuando hablan varios expositores, siempre pone fechas límite para terminar trabajos importantes, algunas veces actúa de manera impulsiva y no domina la lectura de textos académicos en inglés, entonces sí será dado de baja.

d) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, si sabe escribir apuntes en inglés cuando hablan varios expositores, pero no tiene dominio para leer instructivos en inglés, entonces sí será dado de baja.

Como se comentó, para la clase de los estudiantes que sí fueron dados de baja existen otras 6 reglas para terminar de describir a los estudiantes que perteneces a dicha clase. Sin embargo, con estas cuatro reglas se puede clasificar correctamente 45.2% que corresponde a 28 estudiantes de los 62 dados de baja.

Cabe resaltar que de las cuatro reglas anteriores se puede dar una mala lectura, porque se pensaría que si el estudiante vive con su madre, entonces va a ser dado de baja, pero para que esto suceda deben de cumplirse los demás factores de la regla. Lo único que podría decirse a partir de este nodo raíz es que si el estudiante no vive con su madre, entonces tiene una probabilidad de continuar con sus estudios de 4.5% que corresponde a 11 estudiantes de los 243 que no fueron dados de baja.

Por otra parte, para los estudiantes que permanecen en sus estudios existe una regla con soporte de 22.6%, la cual es:

e) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, no sabe escribir apuntes en inglés cuando hablan varios expositores, frecuentemente pone fechas límite para terminar trabajos importantes, no vive con su pareja y el ingreso mensual del padre o tutor es menor a $5 000, entonces no será dado de baja.

Las siguientes reglas en orden de importancia para que un estudiante no sea dado de baja tienen un soporte de 21%, 12.8%, 11.5% y 9.9% respectivamente:

f) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, no sabe escribir apuntes en inglés cuando hablan varios expositores, siempre pone fechas límite para terminar trabajos importantes y nunca actúa de manera impulsiva, entonces no será dado de baja.

g) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa y no cursó geometría en el bachillerato, entonces no será dado de baja.

h) Si el estudiante vive con su madre, no vive con personas distintas a su familia, tiene a lo más una computadora en casa, cursó geometría en el bachillerato, no sabe escribir apuntes en inglés cuando hablan varios expositores, algunas veces pone fechas límite para terminar trabajos importantes, entonces no será dado de baja.

i) Si el estudiante vive con su madre, si vive con personas distintas a su familia, entonces no será dado de baja.

En total se tienen 23 reglas para el grupo de estudiantes que permanecen con sus estudios; sin embargo, con las cinco reglas presentadas se tendría un 77.8% en esa clase que corresponde a 189 estudiantes. Sólo 16 factores aportaron información determinante para saber si un estudiante tiene riesgo de ser dado de baja, que en comparación de los 186 factores iniciales que censa EXANI-II representan una reducción de 91.39%, por lo que sería recomendable cambiar el resto de las preguntas referentes a estos atributos por las que sean similares a los 16 encontrados.

Como eficiencia de clasificación final el algoritmo C4.5 pudo clasificar correctamente 75.1% considerando ambas clases. Se puede observar además que las personas con las que viva el estudiante influyen en gran medida en su desempeño académico, específicamente si vive con su madre, con personas distintas a su familia o con su pareja. También se puede resaltar que existen cinco nodos que se refieren al dominio del inglés como uno de los factores que determina su desempeño académico (ver anexo, figura A1).

3. 1. 1. Relación de los factores encontrados del primer experimento con los obtenidos en el estado del arte

De acuerdo con Pontón (2006) el uso excesivo de la computadora e internet disminuye el desempeño del estudiante. En este caso se ve una relación con la regla *b*).

También es posible relacionar los factores encontrados por K. Bakar (Kamariah *et al.,* 2010) sobre la actitud hacia el aprendizaje y por Nguyen (Nguyen y Griffin, 2010) para el logro académico con el factor sobre la frecuencia de imponerse fechas límite para terminar trabajos importantes.

Otro de los factores relacionados con Oyalinka A. (Adesehinwa y Aremu, 2010), Valero (2010) y Tomul (2009) tiene que ver con las variables familiares, las cuales estarían relacionadas con los factores sobre sí vive con su madre, con personas distintas a su familia o con su pareja. Por último, como señala Valero (2010), el nivel de inglés es otro factor cuando se lee, escribe y habla este idioma.

3. 2. Segundo experimento

Con el objetivo de saber si la información de control escolar del estudiante permite obtener reglas con un mejor soporte se realizó un segundo experimento. El árbol está representado en el anexo, figura A2, el cual permitió aumentar la precisión a 82.6% y mejorar significativamente respecto al primer experimento. El porcentaje de créditos aprobados al sexto semestre es el factor que aporta más información, ahora el nodo raíz.

Al igual que el árbol de decisión del anexo, figura A1, los factores de si vive con su madre y de si vive con personas distintas a su familia vuelven a aparecer en el segundo experimento y se reafirma que estos dos atributos son importantes en el desempeño académico de los estudiantes.

Para el conjunto de estudiantes que sí fueron dados de baja, las reglas más significativas generadas en el segundo experimento son las cuatro siguientes con 29%, 16.1%, 12.9% y 8.1% de precisión respectivamente, las cuales suman 66.1%:

j) Si el estudiante tiene a lo más 38.97% de créditos aprobados, vive con su madre, no vive con personas distintas a su familia, fueron exigentes donde cursó el bachillerato y es hábil para escribir y mandar correos electrónicos, entonces sí será dado de baja.

k) Si el estudiante tiene a lo más 38.97% de créditos aprobados, vive con su madre, no vive con personas distintas a su familia, fueron exigentes donde cursó el bachillerato, es muy hábil para escribir y mandar correos electrónicos, siempre se impone fechas límite para terminar trabajos importantes, sabe escribir notas y cartas cortas con información personal y algunas veces le falta tiempo para concluir actividades que programa, entonces sí será dado de baja.

l) Si el estudiante tiene a lo más 38.97% de créditos aprobados, vive con su madre, no vive con personas distintas a su familia y fueron poco exigentes donde cursó el bachillerato, entonces sí será dado de baja.

m) Si el estudiante tiene a lo más 38.97% de créditos aprobados, vive con su madre, no vive con personas distintas a su familia, fueron exigentes donde cursó el bachillerato, es muy hábil para escribir y mandar correos electrónicos, frecuentemente se impone fechas límite para

terminar trabajos importantes y no sabe escribir en inglés notas o cartas con temas familiares, entonces sí será dado de baja.

En cambio, si nos interesa saber las reglas más significativas para los estudiantes que permanecen en sus estudios, se tendría una sola regla muy significativa con 79% de los casos, la cual es:

n) Si el estudiante tiene más de 38.97% de créditos aprobados (al sexto semestre), entonces no será dado de baja.

Si se considera que al sexto semestre un estudiante debió cursar 60% de los créditos, entonces puede retrasarse hasta dos semestres y tener una alta certidumbre de continuar con sus estudios. Es decir, la trayectoria académica que el estudiante genere en el transcurso de su vida académica es fundamental; al igual que el dominio del inglés, las personas con las que viva el estudiante son significativas, ya que en tres reglas estaría involucrado este factor. Por último, la trayectoria académica que tiene el estudiante antes de ingresar a la universidad es muy importante, pues aparecen dos nodos del árbol de decisión haciendo referencia a su bachillerato.

Conclusiones

Este trabajo propone, a diferencia de la mayoría de los trabajos del estado del arte, una nueva forma de realizar el análisis de las bases de datos de estudiantes mediante minería de datos. En específico, se utilizó el algoritmo C4.5 como la técnica de clasificación, ya que tiene la ventaja, sobre algunas otras, de hacer un análisis completo puesto que es posible trabajar con datos categóricos y con ausencia de información, como sucedió con la analizada.

Cabe resaltar que se contemplaron 206 factores en la base de datos inicial, por lo que esta investigación es una de las que mayor número de factores iniciales contempló. Una vez realizado el análisis, se generó conocimiento potencialmente útil, el cual a diferencia de Valero (2010), nuestro objetivo no es predecir (pues existen otros algoritmos de minería de datos que pudieran ser mejores para al propósito), sino encontrar el conjunto de reglas que mejor explique la deserción y permanencia de estudiantes universitarios.

De acuerdo con la experimentación, es posible decir que existe información valiosa en la base de datos del EXANI-II que permite saber con un 75.1% de certidumbre qué estudiante está o no en riesgo de desertar de sus estudios antes de ingresar a sus estudios de nivel superior. Una vez que haya comenzado, es posible aumentar la información con la base de datos de control escolar, lo cual permite aumentar la precisión del clasificador en 82.6%. Si bien los porcentajes obtenidos tienen una certidumbre aceptable, desde nuestro punto de vista son más relevantes, ya que es posible explicar de manera sencilla (mediante reglas) las fortalezas en su permanencia de estudios y las debilidades en el riesgo de deserción a cada estudiante. Asimismo, hay factores que inciden invariablemente como el de *si el estudiante vive con su madre*.

Además, existe otro factor a considerar para el seguimiento de la trayectoria, que es el número de créditos aprobados. Esto resuelve algunas dudas sobre cómo seguir el desempeño de los estudiantes, que en ocasiones se piensa en el promedio general o en el número de materias aprobadas.

De los 206 factores inicialmente censados por el EXANI-II, sólo 16 aportan información suficiente, por lo que es innecesario preguntar los 172 restantes. En este sentido, sería recomendable hacer preguntas relacionadas con los 16 factores encontrados o muy diferentes a las realizadas actualmente.

Como trabajo a futuro se espera continuar con este estudio hasta saber quiénes de los estudiantes analizados se titularon con el objetivo de explicar la eficiencia terminal.

Análisis prospectivo

Antes de hacer el análisis, este estudio tenía la incertidumbre si la información recabada por el EXANI-II era suficiente para obtener reglas que permitieran modelar a los estudiantes que permanecen o que han desertado de sus estudios. Mediante los arboles de decisión se pudo ver que sí existe información para tal problema. Esto abre un abanico de posibilidades dentro de la minería de datos porque existen varios métodos que pueden aplicarse para mejorar la precisión o la explicación de los patrones encontrados. Por consiguiente, es necesario formar equipos interdisciplinarios de profesores, psicólogos, economistas, ingenieros de cómputo, etcétera, para encontrar qué motivaciones o asociaciones están detrás de las reglas encontradas. Por ejemplo, entre los factores de la regla *g*) aparece que si *el estudiante no cursó geometría en el bachillerato*, entonces no será dado de baja, lo cual puede parecer contradictorio puesto que en la UAP Tianguistenco hay tres carreras de ingeniería y una en seguridad ciudadana; por lo que haría falta revisar si esto se da mayormente en las primeras o en la segunda. Por ello, se tendrían que hacer experimentos separando la base de datos por carrera e incluyendo las bases de datos de las siguientes generaciones. De esta manera, se podría contestar preguntas sobre si la modelación debiera de hacerse por carrera y cuánto cambia el modelo construido por generación de ingreso.

Cabe señalar que el objetivo es utilizar estos patrones para canalizar apoyos específicos a los estudiantes que lo necesitan de manera oportuna (los cuales pueden ir desde apoyos económicos, psicológicos o académicos) y no el de utilizar las reglas obtenidas para determinar qué estudiantes serían aceptados a la universidad. Puesto que con las reglas obtenidas es posible medir gradualmente el riesgo de deserción o permanencia de un estudiante desde su ingreso a la universidad, entonces se podría implementar en sistemas de seguimiento citas tutor-tutorado cuando el estudiante presente alguna de las reglas de deserción.

Bibliografía

Adesehinwa, O. A. y Aremu, A. O. (2010). The relationship among predictors of child, family, school, society and the government and academic achievement of senior secondary school students in Ibadan, Nigeria. *Procedia. Social and Behavioral Sciences*, *5*, 842-849.

ANUIES (Asociacion Nacional de Universidades e Instituciones de Educación Superior) (2005). Disponible en http://www.anuies.mx/servicios/ d_estrategicos/documentos_estrategicos/21/ 2/13.htm

De Vries, W., León Arenas, P., Romero Muñoz, J. F. y Hernández Saldaña, I. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de la educación superior*, *4*(160), 29-50.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P., 1997. *From data mining to knowledge discovery in databases*. Rhode island: AAAI.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (s. f.) The Weka data mining software: an update. *SIGKDD Explorations*, *11*(1), 10-18.

Hassanbeigi, A. y Askari, J. (2010). A study of the most important risk factors of motivational deficiencies in university students. *Procedia. Social and Behavioral Science*, *5*,1972-1976.

Hernández Orallo, J., Ramírez Quintana, M. J. y Ferri Ramírez, C. (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación.

Kamariah, A. B., Ahmad Tarmizi, R., Mahyuddin, R., Habbibahand Wong, S. L., Mohd Ayub, A. F. (2010). Relationships between university students' achievement motivation, attitude and academic performance in Malaysia. *Procedia. Social and Behavioral Sciences*, *2*(2), 4906-4910.

López, B. (2011). Instituto Tecnológico de Nuevo Laredo. Disponible en http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

Lowis, M. y Castley, A. (2008). Factors affecting student progression and achievement: prediction and intervention. A two year study. *Innovations in education and teaching international*, *45*(4), 333-343.

Murillo, J. (2003). Una panorámica de la investigación iberoamericana sobre eficacia escolar. *Revista electrónica iberoamericana sobre calidad, eficacia y cambio en la educación*, *1*(1), 1-14.

Nguyen, C. y Griffin, P. (2010). Factors influencing student achievement in Vietnam. *Procedia. Social and Behavioral Sciences*, *2*(2), 1871-1877.

Pontón, M. C. (2006). Factores que afectan el desempeño de los alumnos mexicanos en edad de educación secundaria. Un estudio dentro de la corriente de eficacia escolar. *Revista elecctrónica Iberoamericana sobre Calidad, eficacia y Cambio en Educación*, *4*(3) 30-53.

Serin, N. B. (2010). Factors affecting the locus of control of the university students. *Procedia. Social and Behavioral Sciences*, *2*(2) 449-452.

Tomul, E. (2009). The relationship between the students' academics achievement and and their socioeconomic level: cross regional comparison. *Procedia. Social and Behavioral Sciences*, *1*(1), 1199-1204.

UAEM (Universidad Autónoma del Estado de México) (2011). *Agenda estadística 2009*. Disponible en http://www.uaemex.mx/tercer informe/3erINFORME_WEB/fscommand/AgEst2009.pdf [Último acceso: 2012 Julio 11].

Valero, S. (2010). *Desarrollo de una herramienta de análisis de datos para predecir deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros*, Puebla: Universidad Popular Autónoma del Estado de Puebla.

Witten, I. y Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Burlington: Elsevier.

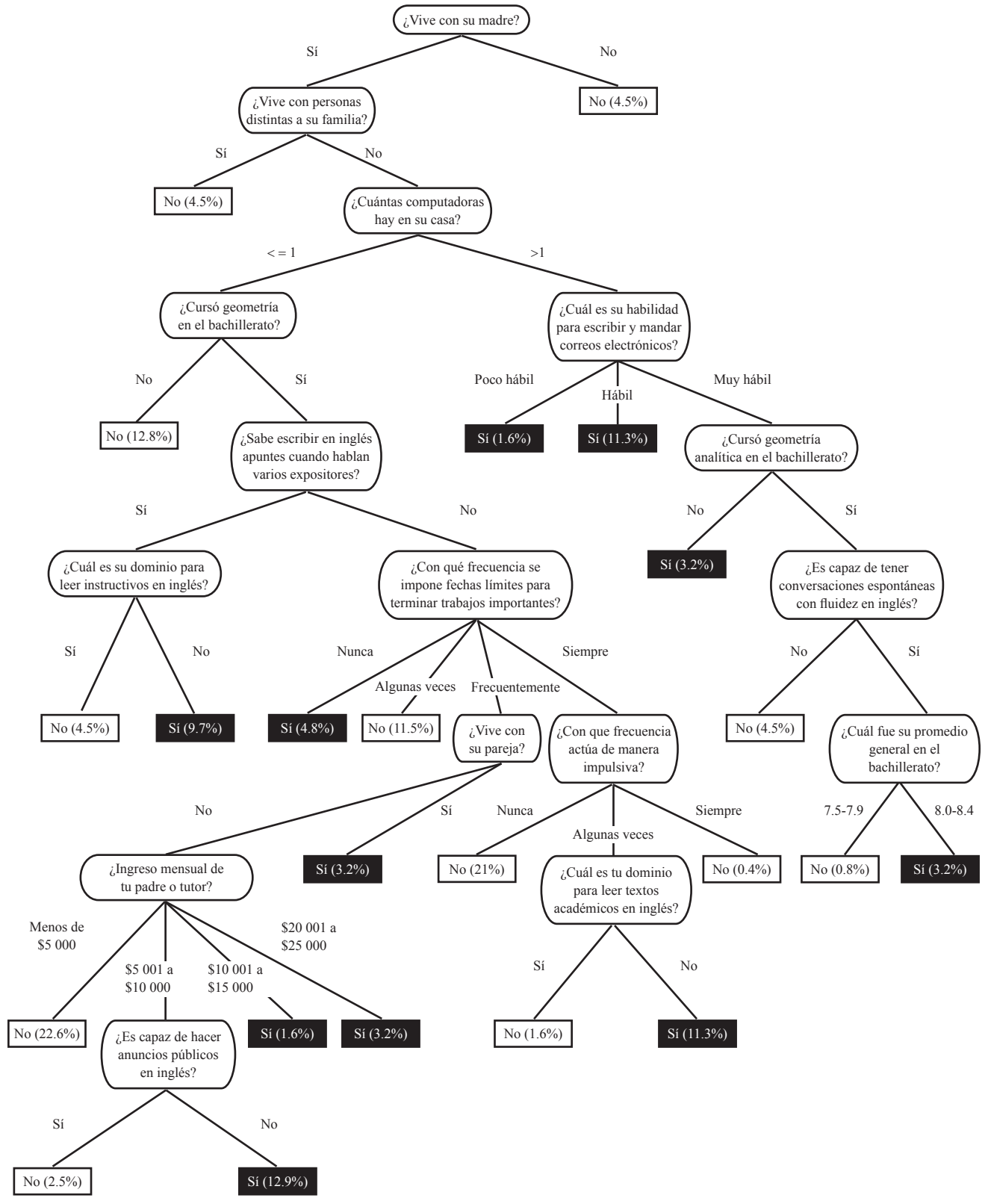
Wu, X., Kumar, V. y Quinlan, J. R. (2008). Top 10 algorithms in data mining. *Knowledge Information Systems*, *14*, 1-37.

Anexos

Tabla A1. Trabajos y factores que inciden en los estudiantes.

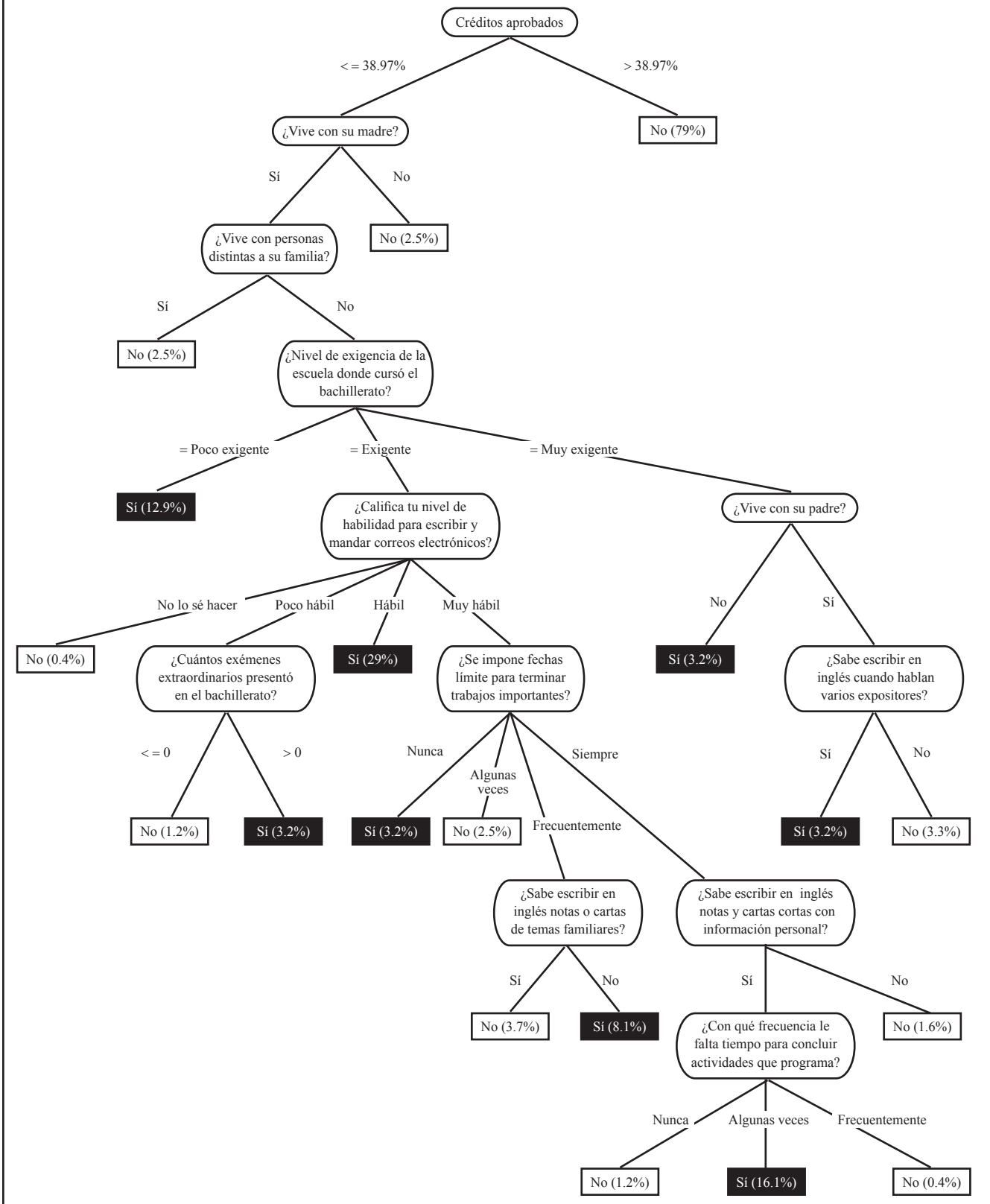
Trabajo	La fuente de datos	Factores encontrados	Modelo de análisis
Kamariah *et al.,* 2010	Cuestionario aplicado a 1 484 estudiantes que considera información demográfica, actitudinal y de motivación al logro	Los resultados mostraron un efecto global significativo en el rendimiento académico de los estudiantes resumidos en: *a*) motivación de los estudiantes a los logros, *b*) actitudes hacia el aprendizaje, *c*) influencia de los compañeros en el aprendizaje y *d*) el género y el grupo étnico al que pertenece el estudiante.	Estudio cuantitativo descriptivo correlacional.
Adesehinwa y Aremu, 2010	Auto-cuestionario que evalúa algunos factores que afectan el desempeño académico	Existe una relación significativa entre los efectos combinados del estudiante que afectan su rendimiento estudiantil como son: *a*) del estudiante, *b*) de la familia, *c*) de la escuela, *d*) de la sociedad y e) del gobierno	Modelo estadístico (análisis de regresión)
Nguye y Griffin, 2010	Muestra nacional de 59 601 estudiantes de Vietnam que evalúa 53 factores en tres niveles: *a*) nivel del estudiante, *b*) nivel de la escuela y *c*) nivel socioeconómico	*a*) Existe un fuerte lazo entre el nivel socioeconómico y el logro académico, *b*) Educación de los padres y *c*) escuelas de tiempo completo.	Modelo estadístico (modelo jerárquico lineal)
Hassanbeigi, A. y J. Askari, 2010	Cuestionario aplicado a 272 estudiantes que evalúa cuatro necesidades básicas	*a*) Necesidades psicológicas básicas, *b*) necesidades psicosociales, *c*) necesidades espirituales, *d*) necesidades educativas.	Modelo estadístico (análisis de media-varianza)
Pontón, 2006	Pruebas estandarizadas como INEE (Instituto Nacional para la Evaluación de la Educación), PISA (Programme for International Student Assessment) y EXANI-I (Examen Nacional de Ingreso a la Educación Media Superior)	Los factores más significantes de las pruebas estandarizadas en cuestión son: *a*) el efecto combinado escuela-maestro en países desarrollados oscila entre 12 y 20%, *b*) el desempeño disminuye conforme aumenta la edad del estudiante, *c*) mujeres con superioridad en lectura o lengua, hombres con superioridad en matemáticas y ciencias, *d*) repetir el año impacta negativamente, *e*) entre mayores sean las expectativas del estudiante sobre su desarrollo, mejor es su desempeño académico y *f*) el uso excesivo de la computadora disminuye el desempeño del estudiante	Modelo estadístico (Modelo jerárquico lineal)
Serin, 2010	Cuestionario de 29 preguntas	Los factores que afectan el locus de control de los estudiantes universitarios son: *a*) el género. Los estudiantes de género masculino tienen un mayor control interno que los estudiantes de género femenino y *b*) nivel socioeconómico. Los estudiantes que se consideran de un nivel socioeconómico elevado tienen mayor control de locus interno que los estudiantes que se consideran con un nivel socioeconómico medio.	Escala de control interno-externo de Rotter (1966).
Lowis y Castley, 2008	Cuestionario con 7 preguntas	*a*) Las mujeres son más consciente de invertir más tiempo en el estudio independiente, *b*) los estudiantes que no están dispuestos a trabajar en equipo no se involucran en actividades de aprendizaje, *c*) los estudiantes requieren más tiempo de contacto con los tutores, *d*) un aspecto clave de rendimiento real fue la necesidad de emplear mucho tiempo en el estudio independiente, *e*) las mujeres tuvieron una apreciación más realista de la necesidad de pasar tiempo en el estudio independiente y *f*) los estudiantes que abandonaron el curso fueron quienes tienden a no trabajar en equipo con otros estudiantes.	Técnica estadística de comparación de medias
Tomul, 2009	Datos de PISA 2006 aplicado a estudiantes de 15 años en Turquía	Las variables familiares tienen mayor efecto en matemáticas y menor efecto en la lectura.	Técnica estadística de regresión lineal simple
Murillo, 2003	Estado del arte de la investigación sobre eficacia escolar en Iberoamérica	Algunos de los factores de eficacia escolar que han considerado en estudios anteriores y que destacan son: Respecto a los factores escolares: *a*) clima escolar, *b*) recursos de la escuela y *c*) trabajo en equipo. Respecto a los factores de aula: *a*) clima del aula, *b*) dotación y calidad del aula y *c*) recursos curriculares. Respecto a los factores asociados al personal docente: *a*) estabilidad, *b*) experiencia, *c*) relación maestro-estudiante y *d*) altas expectativas.	Comparación de trabajos del estado del arte
Valero, 2010	27 factores del examen EXANI-II y de control escolar	*a*) Edad, *b*) ingresos familiares y *c*) nivel de inglés.	C4.5 y K-vecinos más cercanos

Figura A1. **Árbol de decisión considerando la base de datos del EXANI-II.**



Fuente: elaboración propia.

Figura A2. **Árbol de decisión obtenido de la base de datos EXANI-II y control escolar.**



Fuente: elaboración propia.