UNIVERSIDAD DE CÓRDOBA

TESIS DOCTORAL

PREDICCIÓN DEL FRACASO Y EL ABANDONO ESCOLAR MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

Realizada por:

Carlos Márquez Vera

Directores:

Dr. D. Sebastián Ventura Soto

Dr. D. Cristóbal Romero Morales

Fecha: Junio de 2015

TITULO: *Predicción del fracaso y abandono escolar mediante técnica de minería de datos*

AUTOR: *Carlos Márquez Vera*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

La memoria titulada “Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos”, que presenta Carlos Márquez Vera para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado Ingeniería y Tecnología del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, bajo la dirección de los doctores Cristóbal Romero Morales y Sebastián Ventura Soto cumpliendo, en su opinión, los requisitos exigidos a este tipo de trabajos.

Córdoba, Junio de 2015

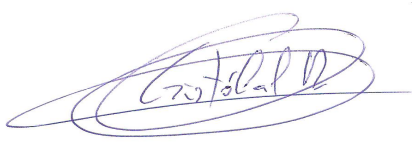
El Doctorando

Fdo: Carlos Márquez Vera

El Director

Fdo: Dr. Sebastián Ventura Soto

El Director



Fdo: Dr. Cristóbal Romero Morales

Esta tesis ha sido parcialmente subvencionada con el Proyecto TIN2014-55252-P del Subprograma de Generación de Conocimiento, Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia.

MINISTERIO DE
ECONOMÍA
Y COMPETITIVIDAD

SECRETARÍA DE ESTADO DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

SECRETARÍA GENERAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN

DIRECCIÓN GENERAL DE INVESTIGACIÓN CIENTÍFICA Y TÉCNICA

SUBDIRECIÓN GENERAL DE PROYECTOS DE INVESTIGACIÓN

AGRADECIMIENTOS

A mis directores de tesis Dr. Cristóbal Romero Morales y Dr. Sebastián Ventura Soto por todo el apoyo que me han brindado, por su disposición y por compartir conmigo sus conocimientos y experiencia.

A mis compañeros del grupo KDIS, especialmente a Marco Antonio Barrón y Alberto Cano quiénes me han ayudado siempre que los he necesitado.

A la Universidad Autónoma de Zacatecas por brindarme el apoyo y las condiciones para realizar mis estudios de doctorado.

A mi familia por apoyarme en todo momento, especialmente a mis hijos Carlos Alberto y Montserrat por su comprensión.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ACRÓNIMOS

AA	Academic Analytics
AEHS	Adaptive Educational Hypermedia System
CENEVAL	Centro Nacional de Evaluación
DM	Data Mining
DSS	Decision Support System
EDM	Educational Data Mining
EWS	Early Warning System
EXANI	Examen Nacional de Ingreso
FN	False Negative
FP	False Positive
G3P	Grammar-Based Genetic Programming
GP	Genetic Programming
ICRM	Interpretable Classification Rule Mining
ITS	Intelligent Tutoring Systems
KD	Knowledge Discovery
LA	Learning Analytics
LMS	Learning Management System

ML	Machine Learning
ROC	Receiver Operating Characteristics
SIAT	Sistema de Alerta Temprana
SMOTE	Synthetic Minority Over-sampling Technique
SNA	Social Network Analysis
TN	True Negative
TP	True Positive
UAPUAZ	Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas

RESUMEN

En esta memoria de tesis se aborda el problema multifactorial del fracaso escolar de los estudiantes. Este problema se presenta en muchas instituciones educativas de todas partes del mundo, especialmente en niveles de educación media o superior, donde hay una cantidad importante de estudiantes que no aprueban sus materias o que abandonan sus estudios. En la actualidad este problema, se ha convertido en una de las prioridades de las instituciones educativas debido a las repercusiones que tiene no sólo desde el punto de vista educativo sino económico. Esta tesis hace una revisión del estado del problema además de proponer varios modelos de predicción a niveles educativos de enseñanza media. Por un lado, se propone una metodología para predecir a los alumnos que se encuentran en riesgo de abandonar o reprobar utilizando diferentes técnicas de Minería de Datos. El objetivo es obtener un modelo de predicción lo más preciso posible, el cual pueda usarse en generaciones futuras para poder reducir la reprobación y el abandono escolar. Por otro lado, se propone, otra metodología también basada en técnicas de Minería de Datos para predecir lo más temprano posible en el periodo escolar a aquéllos que estén en riesgo de suspender o abandonar, es decir, sentar las bases para que se pueda implementar un Sistema de Alerta Temprana, para una vez detectados poder tomar decisiones en cuanto a qué tipo de apoyo o intervención requiere cada uno de ellos para en lo posible impedir el fracaso, o bien, reducirlo y retener a los estudiantes en la escuela. Finalmente, se han realizado diferentes pruebas experimentales con datos reales procedentes de estudiantes de México. Los resultados obtenidos con las metodologías y algoritmos propuestos son buenos y mejoran a las anteriores propuestas con las que se han comparado.

1. INTRODUCCIÓN

En este primer capítulo se introduce al trabajo realizado para concretar esta memoria de tesis. Para ello, primeramente se realiza el planteamiento del problema a resolver, posteriormente se enuncian y describen los objetivos a conseguir a lo largo del trabajo, después se habla sobre las aportaciones que la Minería de Datos hace al campo de la Educación y, finalmente, se comenta el contenido de cada capítulo de esta memoria de tesis.

1.1 PLANTEAMIENTO DEL PROBLEMA

La definición de abandono escolar difiere entre investigadores, pero de cualquier manera si una institución pierde a sus alumnos, se verá reflejado en sus índices de retención. El abandono o deserción escolar de un alumno en un determinado curso puede consistir en la no matriculación de dicho alumno al curso siguiente (Más-Estellés, y otros, 2009). Una definición más genérica es la ausencia definitiva y sin causa justificada del centro escolar por parte del alumno sin haber finalizado la etapa educativa que se esté cursando. Para muchos alumnos abandonar la escuela es el paso final de un largo proceso de desenganche gradual y participación reducida en el currículo formal de la escuela, así como en el co-curriculum y vida social más informal de la escuela (González González, 2006). Por otro lado, la reprobación o la no aprobación de una o más materias se define como un insuficiente rendimiento cuantitativo y/o cualitativo de las potencialidades de un alumno para cubrir los parámetros mínimos establecidos por una institución educativa (Rodallegas Ramos, Torres González, Gaona Couto, Gastelloú Hernández, Lezama Morales, & Valero Orea, 2010). Es también el resultado de un proceso que detiene, limita o no acredita el avance del alumno en su vida académica (Corral Verdugo & Diáz Núñez, 2009).

En la mayoría de las instituciones educativas de nivel medio y superior de todas partes del mundo, los índices de reprobación y abandono presentan unos valores muy altos al compararlos con los resultados de la educación básica, donde prácticamente todos los estudiantes aprueban. De hecho algunos de los indicadores más importantes que reflejan la calidad que tiene una institución educativa son junto con el número de estudiantes matriculados, los mencionados índices de reprobación y abandono (Zhang, Oussena, Clark, & Kim, 2010). Y con el objetivo de cumplir con su cometido principal de brindar una buena formación académica, en la actualidad las instituciones educativas están muy interesadas en retener a sus estudiantes, es decir, en que aprueben, vayan avanzando en su formación y no abandonen. A pesar de que hay políticas y programas que se han venido implementando para evitar el fracaso escolar de los estudiantes, disminuir la reprobación y el abandono no es una tarea sencilla, ya que es un problema multifactorial al que incluso se le ha denominado "el problema de las mil causas" (Magaña Hernández, 2002).

Por otro lado, la Minería de Datos (*Data Mining*, DM) es un área multidisciplinar que se está usando exitosamente para resolver problemas en muchos campos como los negocios, la ciencia e ingeniería, la salud, la educación, los juegos, etc. DM, se define como un proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y deseablemente entendibles a partir de los datos. En DM confluyen diversos paradigmas de la computación tales como la construcción de árboles de decisión, la generación de reglas de inducción, las redes neuronales, el aprendizaje basado en instancias, los métodos bayesianos o los algoritmos estadísticos (Witten, Frank, & Hall, 2011), y desde hace relativamente poco tiempo también se han utilizado algoritmos de soft-computing como los algoritmos evolutivos. Un caso concreto de aplicación de DM es en el campo de la educación, donde se le denomina Minería de Datos Educativa (*Educational Data Mining*, EDM) (Romero & Ventura, 2007) (Romero & Ventura, 2013). Actualmente se está aplicando para tratar problemas en los sistemas tradicionales de educación para predecir el rendimiento académico de los estudiantes, en los cursos basados en la web, en los sistemas de gestión para el aprendizaje de contenidos, en los sistemas inteligentes de aprendizaje, etc. Entre las diferentes tareas que pueden realizarse con DM están la descripción, la predicción, la segmentación y la asociación. Particularmente la predicción puede realizarse por medio de la clasificación, la cual es una de las actividades que más frecuentemente realiza el ser humano en su vida cotidiana. La clasificación ocurre cuando es necesario asignar un objeto a un grupo predefinido o clase, lo cual se decide basándose en un determinado número de atributos que se observan en el objeto (Zhang, Oussena, Clark, & Kim, 2010). El objetivo de la clasificación es inducir un modelo para predecir a qué clase pertenece un objeto.

Gracias al gran desarrollo de Internet, la enseñanza a distancia y los equipos de informática es cada vez más común que las instituciones educativas tengan suficiente información de sus estudiantes. Generalmente pueden disponer fácilmente de ella, pero no procesarla a la par que se genera u obtiene, y es seguro que se puede encontrar en esos datos "conocimiento oculto" potencialmente útil. Esta gran cantidad de información puede verse como una mina de oro con datos acerca de los estudiantes, tanto de los que aprueban como de los que no aprueban o abandonan. Un problema que puede darse en los datos, es que la mayoría de los estudiantes aprueben y sólo la minoría fracasen (reprueben o abandonen) dando lugar a un conjunto de datos desbalanceados. El tratar de realizar una clasificación en los conjuntos de datos

desbalanceados y obtener buenos resultados no es una tarea sencilla debido a que los algoritmos de clasificación tienden a ignorar los elementos que pertenecen a la clase minoritaria. Además, es importante indicar que en entornos educativos clasificar correctamente a los elementos de esta clase minoritaria es lo más importante.

Para resolver este problema, en esta memoria de tesis, se van a proponer dos metodologías: una de ellas tiene por objetivo detectar con la máxima precisión y otra lo más temprano posible a los estudiantes que están en riesgo. Ambas metodologías se han probado experimentalmente con información real de estudiantes de una institución de educación media superior de México, concretamente el Programa II de la Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas. El poder predecir de manera suficientemente confiable a quienes están en riesgo potencial de reprobar o abandonar, permite al personal directivo de la institución el estar en condiciones de tomar decisiones para implementar acciones que puedan apoyar de distintas maneras a los alumnos y así tratar de disminuir el fracaso escolar.

1.2 OBJETIVOS

El objetivo principal de este trabajo es obtener un modelo capaz de predecir los estudiantes que potencialmente presentan un mayor riesgo de fracasar académicamente, ya sea por reprobar o abandonar. Para buscar ese modelo se va a utilizar técnicas de DM y particularmente algoritmos de clasificación. Además en este trabajo se hace uso de información real de jóvenes estudiantes de una institución educativa de nivel de enseñanza medio-superior de México, dentro de un sistema de educación tradicional, es decir, presencial. Finalmente, para conseguir este objetivo general mencionado, se deben de alcanzar otros sub-objetivos particulares, los cuales se describen a continuación:

- Desarrollar e implementar una metodología para la predicción de la reprobación y el abandono de los estudiantes.
- Probar que los algoritmos de clasificación pueden usarse para obtener una buena predicción del rendimiento académico de los estudiantes al final del periodo escolar.

- Realizar un análisis del rendimiento de diferentes algoritmos de clasificación al ser aplicados a distintos conjuntos de datos.
- Reducir la dimensión de un conjunto de datos a partir de la técnica de selección de mejores atributos. Verificar y comparar la precisión de los algoritmos de clasificación con un conjunto de datos antes y después de ser reducido.
- Obtener buenos resultados de clasificación de un conjunto de datos desbalanceado, especialmente obtener una buena clasificación de la clase minoritaria, la cual es la de mayor interés de este trabajo.
- Obtener un modelo para la predicción de los estudiantes en riesgo de fracasar que pueda ser utilizado en un SIAT (Sistema de Alerta Temprana) o EWS (*Early Warning System*).

1.3 CONTRIBUCIONES

La principal contribución realizada en esta memoria de tesis es la propuesta de dos metodologías generales basadas en técnicas de clasificación para la predicción del abandono y fracaso escolar. La primera metodología tiene la finalidad de predecir a los estudiantes en riesgo de reprobar o abandonar sus estudios con la mayor exactitud posible. La segunda metodología tiene por objetivo predecir lo más temprano posible a los estudiantes en riesgo de reprobar o abandonar sus estudios. Ambas son generales y pueden ser utilizadas como guía en cualquier institución educativa que tenga intenciones similares a las que se han planteado en este trabajo. Además también se propone aplicar una etapa específica de pre-procesado de los datos para la selección de los mejores atributos, lo cual permita una optimización del proceso al reducir la enorme cantidad de atributos disponibles. Finalmente, se va a utilizar específicamente información procedente de una institución educativa de nivel medio superior. A destacar que la gran mayoría de los trabajos relacionados se han realizado en otros niveles y modalidades educativas.

1.4 CONTENIDO DEL DOCUMENTO

A continuación se describen brevemente los distintos capítulos en los que está dividida (ver Figura 1.1) la memoria de esta tesis:

- El capítulo *Introducción* presenta una visión general de este trabajo. Se realiza el planteamiento del problema a resolver y su justificación e importancia, además se citan los objetivos que se pretenden conseguir y las contribuciones que se hacen al campo de EDM.

- El capítulo *Antecedentes* revisa los problemas de reprobación y abandono escolar de los estudiantes, así como cuáles son las causas más comunes que provocan estos problemas. También se introduce la utilización de EDM en la predicción de los alumnos en riesgo de fracasar y finalmente se aborda cómo se han implementado los SIAT en entornos educativos.

- El capítulo *Predicción del fracaso escolar de los estudiantes utilizando diferentes técnicas de Minería de Datos* propone una metodología para abordar el problema de predicción de los estudiantes en riesgo. Se describen varios experimentos realizados con diferentes técnicas de minería de datos, con la finalidad de obtener una buena clasificación de los alumnos que fracasan.

- El capítulo *Predicción temprana del fracaso escolar usando Minería de Datos* establece una metodología para poder predecir confiablemente y lo más temprano posible a los estudiantes en riesgo de abandonar o reprobar. Se describen los diferentes experimentos realizados y los resultados obtenidos.

- El capítulo *Conclusiones y Trabajo futuro* resume lo que se ha conseguido con esta memoria de tesis y se plantean algunas posibilidades de mejoras y trabajos en el futuro en nuevas líneas abiertas a partir de este trabajo.

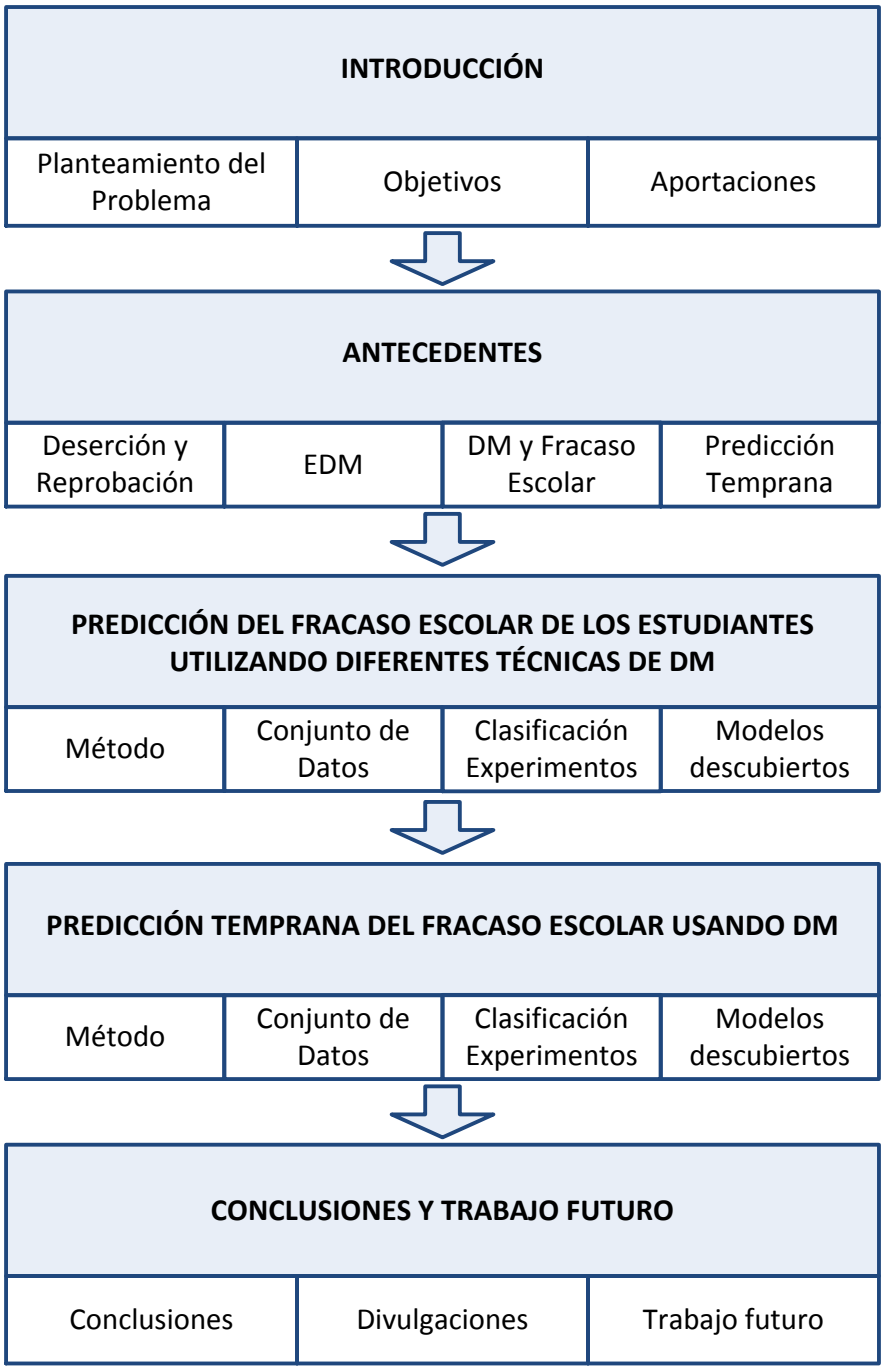


Figura 1.1 Esquema general de la tesis.

2. ANTECEDENTES

En este capítulo se introducen los problemas de abandono o deserción y el fracaso escolar de los estudiantes. Se enumeran las variables que se han encontrado en estudios anteriores como aquellas que más influyen en los estudiantes para que no aprueben, para que deserten o que decidan abandonar sus estudios. También se profundiza en la aplicación de técnicas de DM en el campo de la educación o EDM, principalmente las técnicas de clasificación y como son utilizadas para predecir la reprobación y el abandono escolar de los estudiantes.

2.1 INTRODUCCIÓN AL PROBLEMA DE ABANDONO Y FRACASO ESCOLAR DE LOS ESTUDIANTES

La educación es un factor estratégico para el desarrollo y bienestar de cualquier sociedad, si la población no tiene acceso a la educación entonces su nivel educativo es deficiente y su crecimiento y desarrollo se ven limitados. Por esto las posibles soluciones a algunos de los problemas que se presentan en los estudiantes como el abandono o deserción y la reprobación, entre otros, son temas vigentes y actuales que necesitan el aporte de todos los actores que participan en el proceso educativo. En los últimos años ha surgido en muchos países una creciente preocupación ante el problema del fracaso escolar de los estudiantes y por determinar los múltiples factores que pueden influir en él (Álvarez Aldaco, 2009). De hecho, la mayoría de los trabajos publicados que tratan sobre este tema están enfocados en determinar cuáles son los factores que más afectan al rendimiento académico de los estudiantes. Estos estudios se han realizado en los diferentes niveles educativos: en la educación básica, media y superior (Araque, Roldán, & Salguero, 2009).

Uno de los objetivos de cualquier institución educativa es brindar las condiciones para que sus estudiantes obtengan la mejor formación académica posible y que puedan desarrollar sus capacidades. Evidentemente esto puede llegar a alcanzarse en los estudiantes que persisten, no en aquéllos que fracasan (desertan, abandonan o no aprueban). El abandono escolar es uno de los indicadores más importantes de la calidad de los sistemas educativos ya que muestra si hay fallos en el proceso de orientación, transición, adaptación y promoción de los estudiantes. Por esta razón, en el marco del espacio europeo de la educación superior muchas universidades están tomando en cuenta en sus planes estratégicos, como primer objetivo, disminuir el abandono escolar (Araque, Roldán, & Salguero, 2009). Las instituciones educativas que presentan una baja tasa de abandono o deserción, obtienen además beneficios como el recibir apoyos institucionales extraordinarios (Parker, 2003). En España por ejemplo, las universidades reciben recursos extraordinarios del gobierno a través de lo que se conoce como contrato-programa, en éste se recogen un conjunto de acciones a realizar por parte de las universidades que condicionan la financiación en función de las siguientes variables: calidad de la docencia universitaria, investigación, control financiero a través de auditorías periódicas de gestión, financieras y de seguimiento del contrato-programa y mejora del nivel de

información. Para ello se realiza un seguimiento de las tasas de graduación, de abandono, de eficiencia, de éxito, de rendimiento y la duración media de los estudios (Esparrells, 2004).

Por otro lado, en México, la deserción, el rezago estudiantil y los bajos índices de eficiencia terminal, se encuentran entre los problemas más complejos y frecuentes de las instituciones de educación superior del país, y en la actualidad son reconocidos prácticamente por todas ellas (Valero Orea, 2009). Actualmente existen programas o proyectos a nivel nacional en los que las instituciones educativas pueden participar para obtener recursos extraordinarios, pero éstos deben ser empleados en rubros específicos que la Secretaria de Educación Pública designa, entre los cuales se encuentra prevenir y reducir el abandono escolar. Uno de estos programas federales es el PAAGES (Proyecto para el Avance en la Autonomía de la Gestión Escolar) (Subsecretaria de Educación Media Superior SEMS, 2014).

Según la Organización para la Cooperación y Desarrollo Económicos (OCDE), tan sólo un 25% de los jóvenes logran terminar la escolaridad obligatoria y mantienen la posibilidad de incorporarse a una licenciatura o posgrado; un 26% en edad de graduación concluye una licenciatura y sólo el 18% cuenta con certificado de licenciatura y posgrado (Álvarez Aldaco, 2009). La mayoría de las instituciones educativas han hecho algún tipo de esfuerzo por disminuir estos índices realizando y estableciendo programas de tutorías, asesorías, congresos, talleres y otros tipos de eventos, para que los alumnos se involucren directamente y aumente su compromiso. Sin embargo, muchos de estos esfuerzos no han sido suficientes y el fenómeno se sigue repitiendo constantemente (Valero Orea, 2009). El mayor impacto de la deserción, reprobación y repetición se acentúa en los primeros semestres, ya sea en el nivel medio o bachillerato y superior o universitario (Álvarez Aldaco, 2009). Muchos profesores universitarios, particularmente aquéllos que imparten clases en los primeros semestres, tienen la sensación de un elevado fracaso escolar, algo que se apoya en las estadísticas que en los distintos centros se elaboran. Este elevado fracaso tiene unas connotaciones en términos sociales, económicos y humanos que la sociedad no debe ignorar (Más-Estellés, y otros, 2009).

Siempre hay un conjunto de estudiantes en riesgo que podrían ser salvados, es decir, estudiantes que pueden tener éxito pero necesitan una atención especial o individual específica. Detectar a todos los estudiantes con riesgo en una etapa temprana es

fundamental para retenerlos y evitar la reprobación y/o el abandono. Esto permite al departamento institucional correspondiente orientar su esfuerzo donde más se necesita. Sin embargo, la predicción del abandono o deserción es una importante y cambiante tarea (Dekker, Pechenizkiy, & Vleeshouwers, 2009). Un estudio realizado en 1997 (Braxton, Jhonson, & Shaw-Sullivan, 1997) indica que la falta de persistencia en los estudiantes es un problema tan complejo que se constituye en un rompecabezas. El fracaso escolar también es conocido como "el problema de las mil causas" (Magaña Hernández, 2002). En la actualidad hay un creciente interés en los factores que predicen el rendimiento de los estudiantes, probablemente más acentuado en los ambientes de educación a distancia.

Tinto (Tinto, 1975) ha escrito ampliamente sobre el abandono y la deserción escolar y ha enfatizado en la educación superior. En sus trabajos comenta que este acto puede tener significados múltiples y diferentes para los que están implicados o afectados por este comportamiento. Es diferente la visión de la institución educativa a la de los estudiantes. Para un funcionario universitario, la deserción significa un fracaso del alumno en cuanto a completar satisfactoriamente un programa de estudios, mientras que para éste no necesariamente tiene el mismo significado, puede ser un paso positivo hacia la consecución de una meta, y es que no todos tienen como finalidad completar un programa de estudios. Por ejemplo, para algunas personas completar un programa de estudios puede no ser un fin deseable, sino una corta asistencia a la universidad sin necesariamente terminar una carrera, puede ser suficiente para lograr sus metas. Se presentan muchos casos: no solamente desertan aquellos que presentan un bajo rendimiento académico, hay algunos que van consiguiendo buenos resultados académicos y sin embargo abandonan sus estudios, algunos asisten temporalmente a la universidad sólo para conseguir algún tipo de certificación que requieran en su empleo, otros pueden cambiar de institución educativa por interés o por necesidad, otros pueden cambiar de carrera, etc. A este tipo de comportamientos no se les puede catalogar como abandono o deserción en su connotación de fracaso. Se reconoce que alcanzar el éxito depende en buena medida del interés, motivación y habilidades que tengan los jóvenes, en muchos de los casos el abandono no se produce por falta de capacidad. Por supuesto, hay muchos aspectos que influyen, entre ellos están la falta de interés o compromiso y la ausencia de habilidades sociales que impiden que los estudiantes puedan integrarse adecuadamente al entorno de la universidad. Existe gran cantidad de trabajos sobre

las diferentes causas que pueden llevar a la deserción de los distintos grupos de estudiantes de una institución educativa. Por ejemplo, los estudiantes que pertenecen a minorías o que tienen antecedentes desventajosos, pueden tener más problemas para relacionarse y establecer vínculos académicos y sociales con la comunidad universitaria, lo cual puede ser un factor importante para tomar la decisión de abandonar. En cuanto al instante en el que se presenta la deserción, es más frecuente en el primer semestre, aunque también cambia a medida que se avanza en la carrera. Hay una etapa crítica que ocurre durante las primeras seis semanas desde el inicio del primer semestre universitario, que es el periodo de transición del nivel medio al nivel superior de educación. Esto se relaciona con lo que ya se ha mencionado anteriormente, porque también en este semestre los estudiantes todavía no se han integrado al nuevo entorno educativo y no han establecido relaciones sociales en la comunidad universitaria a diferencia de lo que ocurre en semestres más avanzados. Sin embargo, desde el punto de vista institucional no hay diferencia, cualquiera que sea la razón por la que algún estudiante abandona, todos son desertores y estos alumnos desestabilizan a cualquier institución sin importar su carácter, ya sea público o privado. Por las instituciones educativas desean conservar a todos los estudiantes que ingresan. Es evidente que en muchos casos la institución no puede hacer mucho, pero en otros sí y para ello se debe esclarecer cuáles son los motivos principales para que los estudiantes abandonen y así poder impulsar las más adecuadas políticas institucionales para aumentar la retención. Las instituciones desean retener a aquellos estudiantes que tienen el interés en permanecer en la universidad, pero que encuentran dificultades para satisfacer las exigencias académicas o que tienen dificultades de integración social y académica con la comunidad universitaria. Todo estudiante que abandone la Universidad puede categorizarse como desertor, pero no todas las deserciones merecen acciones de la institución para tratar de retenerlos y ninguna universidad puede resolver todos los casos de abandono.

Finalmente, indicar que el fracaso escolar es un término que plantea diversos interrogantes: se puede afirmar que se trata de un fenómeno estrechamente ligado y producido por la escuela, que sólo pueden fracasar aquellos estudiantes que asisten a ella; quienes ni siquiera cuentan con esa oportunidad están libres de fracasar. La privación que sufren de educación es una muestra más de situaciones que corresponden a la redistribución desigual e injusta de acceso y disfrute de bienes

básicos, los que todavía son negados por la sociedad de la opulencia y el bienestar a los seres más indefensos. En países como España y otros, donde se garantiza el derecho de acceso y permanencia en la educación durante años, muchos niños y jóvenes sufren la paradoja de que la misma institución, que ha sido pensada y dispuesta para ayudarles a lograr los aprendizajes considerados indispensables, es la misma que fabrica, sanciona y certifica sus fracasos, su reprobación, su exclusión o abandono (Escudero Muñoz, González González, & Martínez Domínguez, 2009).

2.1.1. FACTORES QUE INFLUYEN EN EL RENDIMIENTO

Existen una gran cantidad de factores que pueden influir en el rendimiento académico de los estudiantes y por ello se le denomina "el problema de las mil causas" (Magaña Hernández, 2002). A continuación se citan los principales factores que aparecen en la bibliografía, que se han agrupado en función del nivel educativo, aunque algunas son generales, hay otras que influyen más en una etapa particular de la educación. Es de destacar que en el nivel superior de educación en sus distintas modalidades es donde más investigación y publicaciones en el tema se han realizado, posteriormente en el nivel básico y finalmente en el nivel medio.

En la Educación Básica los factores que afectan el rendimiento académico se agrupan normalmente en tres categorías:

- **Del Alumno.** Los principales factores o problemas del alumno específicos en el aprendizaje de niños son: el trastorno para realizar la lectura que imposibilita su correcta comprensión (Dislexia), la pérdida de la capacidad de leer cuando ya fue adquirida previamente (Alexia), la dificultad que impide dominar y dirigir el lápiz para escribir de forma legible y ordenada (Digrafía), la pérdida de la destreza en la escritura (Agrafía), la dificultad para escribir las palabras de manera ortográficamente correcta (Disortografía), la dificultad en el proceso de aprendizaje del cálculo (Discalculia), la dificultad para realizar cálculos (Acalculia), la pérdida de la capacidad de llevar a cabo movimientos de propósitos aprendidos y familiares a pesar de tener la capacidad física (Apraxia), el trastorno por déficit de atención con o sin hiperactividad (TDA), el déficit intelectual en el límite de la normalidad, los problemas

neurológicos, las enfermedades crónicas o la discapacidad física, el déficit sensorial auditivo y/o visual, las enfermedades carenciales como la malnutrición, la ferropenia o las alteraciones tiroideas que provocan apatía y/o somnolencia, la rinitis crónica, las adicciones a la televisión a los videojuegos o a la computadora (Magaña Hernández, 2002).

- **Socio-Familiares**. Respecto a los problemas derivados del entorno Socio-Familiar, los más habituales suelen ser: que en la familia o en el hogar del alumno exista la presencia de muchos niños (cinco o más), que los padres o tutores sean analfabetas, que los padres de los niños sean menores de 20 años, la ausencia de material educativo en casa, que el niño no viva con sus padres (ambos), que no tenga apoyo académico en casa, que tenga que involucrarse apoyando en el trabajo en el hogar, las condiciones adversas de la vivienda, su medio de transporte a la escuela, el bajo ingreso familiar, el acceso limitado a los servicios básicos, la ocupación del padre y la madre y la baja escolaridad de los padres (CIET, 1995), (Silas-Casillas, 2009).

- **Escolares**. Respecto a los factores adversos que involucran a la escuela, a los profesores y a los alumnos suelen ser: que los profesores no tengan una plaza permanente (pueden ser aspirantes o interinos), que la escuela cuente con más de 500 alumnos, que los niños no tuvieron formación pre-escolar, que la escuela sea muy pequeña y cuente con un solo docente y no tenga aulas diferenciadas, que la escuela carezca de áreas recreativas, la poca disponibilidad de libros de texto, la falta de equipamiento e infraestructura en la escuela, la poca formación de los profesores, la calidad de la interacción entre profesor-alumno y otros factores asociados al desarrollo de un adecuado clima pedagógico en el aula (CIET, 1995) (Silas-Casillas, 2009) (Koedel, 2008).

En la Educación Media los factores son más típicos de los adolescentes, como pasar mucho tiempo en distracciones como la televisión, los videojuegos, la computadora y el internet, además se presenta el abuso en el consumo de drogas y alcohol. Otras que también afectan son la actitud de violencia, la promiscuidad, el embarazo adolescente, los problemas de conducta, el grupo de amigos y el poco interés que se tiene por estudiar (Espíndola & León, 2002). Para muchos expertos ningún factor es

tan significativo para el rendimiento escolar de los estudiantes como el ambiente familiar, aspecto relacionado con el número total de vástagos en el hogar y el orden de nacimiento que ocupa cada uno de ellos, a mayor cantidad de hermanos se da mayor proporción de fracasos, con el nivel educativo de los padres, con la actitud orientadora de los padres en cuanto al trabajo escolar, factor que también influye en la formación de valores culturales (Araque, Roldán, & Salguero, 2009). El bajo rendimiento escolar (reprobación) está en función (en la mayoría de los casos) de causas de origen psicosocial, incluida la familia y considerando los problemas que se viven dentro de ella como son los conflictos maritales, divorcios, abandono del hogar de alguno de los padres, la falta de atención y motivación de los padres a los hijos y las mismas relaciones que existen entre ellos (Espíndola & León, 2002). No se pueden dejar de lado la situación socioeconómica y el contexto familiar de los niños y jóvenes como fuentes principales de diversos hechos que pueden facilitar directa o indirectamente el abandono escolar como las condiciones de pobreza, de marginalidad y la adscripción laboral temprana (Espíndola & León, 2002) (Chanyoung & Orazem, 2010). Existen circunstancias que pueden alterar el equilibrio afectivo y perjudicar el rendimiento escolar, entre ellas están las situaciones especiales como la muerte o enfermedad de uno de los progenitores o de alguien muy cercano al estudiante, los estilos educativos de los padres como la severidad excesiva, la disciplina extrema, el exceso de perfección y de protección a los hijos (Espíndola & León, 2002). Por último, hay factores del ámbito normativo de las instituciones educativas que influyen en el rezago académico y en la deserción escolar, como los requisitos de ingreso, la seriación de materias, el número de oportunidades para cursar la misma materia, el número permitido de asignaturas reprobadas, los tipos de exámenes y el número de ocasiones que se pueden presentar, el plazo para concluir los estudios y el autoritarismo docente (Álvarez Aldaco, 2009) (Espíndola & León, 2002). Otros factores son el exceso de estudiantes en las aulas particularmente cuando el número es mayor de 25 y el papel que juega el profesor (Espíndola & León, 2002).

En la Educación Superior es donde más investigación en el tema se ha realizado y principalmente en la modalidad “a distancia” o “en línea”, ya que por sus propias características se puede disponer de información de los estudiantes más fácilmente que en la educación tradicional. En este nivel educativo, a diferencia de los anteriores y debido a que el rango de edades de los estudiantes es muy amplio, el número de factores o variables que afectan el desempeño académico es mayor. De forma general

se puede decir que las variables que son consideradas en las diferentes investigaciones son características personales como la edad, el origen étnico, género, estrato social, ciudadanía, expectativas que se tienen sobre la educación, la importancia que se le atribuye al estudio, el estado civil, si tiene hijos, problemas de salud, el medio de transporte a la escuela y si trabaja (Quadril & Kalyankar, 2010) (Xenos, Pierrakeas, & Pintelas, 2002) (Romero & Ventura, 2007) (Levy, 2007) (Inan, Yukselturk, & Grant, 2006) (Romero & Ventura, 2010) (Más-Estellés, y otros, 2009). Las variables que están relacionadas con el entorno socio-familiar son el tipo de relación del estudiante con los integrantes de su familia y el apoyo afectivo, los problemas que se presentan en el seno familiar, el ingreso familiar, el estrato social, nivel cultural de la familia, el nivel educativo de los padres, el número de amigos (Espíndola & León, 2002) (Levy, 2007) (Inan, Yukselturk, & Grant, 2006). Las variables que están relacionadas con la universidad y su entorno son el promedio de calificaciones en el bachillerato, la calificación en algunas asignaturas particulares en el bachillerato como Matemáticas y la relacionada con el estudio del idioma o lengua, el tipo de universidad, el resultado de algunas pruebas como el SAT (*Scholastic Aptitud Test*) o el ALOC (*Academic Locus of Control*), otras que miden el coeficiente intelectual y los intereses vocacionales, si cuenta con beca (el tipo de ella y su monto), el estilo de aprendizaje, el tipo de evaluación que realizan los profesores, el número de asignaturas cursadas, competencia en el uso de la computadora, el tiempo dedicado a estudiar, las técnicas empleadas para estudiar, la organización de la institución, su normatividad e infraestructura, medios y recursos para la enseñanza y el aprendizaje, la relación con el profesorado, exceso de carga curricular, carencia de actividades extracurriculares, la relación de los estudiantes con la institución, la falta de preparación, si el estudiante vive en el campus, la suspensión de los estudios por uno o más periodos escolares y la inscripción de manera parcial. En algunos estudios se encontró un número muy elevado de variables que pueden explicar el abandono y la persistencia, en uno de ellos se encontraron 146 posibles variables. Por otro lado, se identificaron 192 variables del entorno que tienen influencia en el éxito de los estudiantes y fueron organizadas en 8 categorías: características institucionales, estudiantes con características similares, características de los profesores, currículo, ayuda financiera, campo principal de preferencia, lugar de residencia y actividades de participación de los estudiantes (Xenos, Pierrakeas, & Pintelas, 2002) (Romero & Ventura, 2007) (Massa & Puliafito, 1999) (Giménez, 2005) (Romero & Ventura, 2010) (Kotsiantis S. B., 2009)

(Inan, Yukselturk, & Grant, 2006) (Monge Reyes & Martínez Godínez, 2006) (Magaña Hernández, 2002) (Más-Estellés, y otros, 2009) (Reyes-Seánez, 2006).

A continuación, a modo de resumen se presentan y agrupan en distintas categorías todas las variables o factores que más se han citado en la bibliografía (ver Tabla 2.1) como aquellas que pueden influir en los estudiantes para que no acrediten sus estudios o los abandonen.

Tabla 2.1 Variables o factores que influyen en los estudiantes para que reprueben o abandonen.

Categorías	Variables/Factores
Personales	Edad, género, raza, estado de ciudadanía, personalidad, importancia que le da a su educación, tiempo dedicado a estudiar, estado civil, haber reprobado o abandonado la escuela previamente.
Académicas	Calificación del examen de admisión, promedio de calificaciones previo al ingreso, calificación en Matemáticas, calificación en asignatura referente a la lengua o idioma de origen, coeficiente intelectual.
Físicas	Enfermedades, discapacidad.
Económicas	Bajo nivel socioeconómico o de ingreso familiar, contar con una beca.
Familiares	Problemas familiares, bajo apoyo afectivo de los padres al estudiante, deficiente nivel cultural de la familia, número de hermanos del estudiante, orden de nacimiento entre los hermanos, nivel de educación de los padres, actitud orientadora de los padres hacia el trabajo escolar, ausencia de material de apoyo a la educación en casa.
Sociales	Estrato social, marginación, número de amigos en el aula.
Institucionales	Nivel de satisfacción del estudiante con la institución, requisitos de ingreso, número de oportunidades para aprobar una asignatura, número permitido de asignaturas reprobadas, tipos de exámenes y número de ocasiones que se pueden presentar, plazo para concluir los estudios, grupos con más de 45 alumnos por aula, maestros aspirantes o interinos, baja calidad académica de los profesores, escuelas con población estudiantil muy grande, infraestructura en malas condiciones o insuficiente, una gran carga académica para los estudiantes.

Pedagógicas	Control de Locus académico (¿qué controla el desempeño académico del estudiante?, ¿dónde está?), tipo de carrera a estudiar, estilo de aprendizaje, problemas específicos de aprendizaje, carencia de métodos efectivos para el aprendizaje.
Laborales	Tiempo semanal dedicado al empleo.
Adicciones	Exceso de tiempo dedicado a la Televisión, videojuegos o computadora y el consumo de alcohol y/o drogas.
Otras	Retrasar el ingreso a la escuela, permanecer un periodo de tiempo grande sin estudiar.

2.2 MINERÍA DE DATOS APLICADA A LA EDUCACIÓN

La minería de datos aplicada a la educación o EDM es una disciplina emergente que desarrolla y aplica métodos de minería de datos o DM sobre datos que vienen de entornos educativos, y los usa para entender mejor a los estudiantes y los entornos en los que aprenden (Romero & Ventura, 2013). Además del término EDM existen también otros términos similares que son empleados por áreas muy afines o relacionadas con EDM como son:

- **Análisis del aprendizaje** o **Learning Analytics (LA)** consiste en la medida, recolección, análisis e informe de datos sobre estudiantes/aprendices y su contexto, con el propósito de comprender y optimizar el aprendizaje y el entorno donde ocurre.
- **Análisis académico** o **Academic Analytics (AA)** consiste en la aplicación de técnicas estadísticas y de minería a datos institucionales para producir inteligencia de empresa y soluciones a universidades y administradores.

En la actualidad las instituciones educativas, gracias al empleo masivo de los equipos de informática, se cuenta con enormes cantidades de información disponible sobre diferentes áreas del proceso educativo. Esta información puede ser desde las notas de los estudiantes, los hábitos de estudio, tareas, asistencia, comportamiento, asignaturas, el uso de recursos en línea, los profesores, la infraestructura escolar, bibliografía, etc. Toda esta información es una auténtica mina de oro que puede proporcionarnos conocimiento potencialmente útil por ser descubierto. EDM está

interesada en el desarrollo de métodos para explorar los datos en que se dan en el proceso educativo y en transformarlos para entender y atender mejor a los estudiantes y para hacer los ajustes necesarios para que éstos puedan aprender en mejores condiciones (Romero & Ventura, 2010) (Heiner, Baker, & Yacef, 2006). Desde un punto de vista práctico, EDM permite descubrir conocimiento basado en los estudiantes, usando información para ayudar a evaluar o validar los sistemas educativos, para mejorar algunos aspectos de la calidad de la educación y para sentar las bases para obtener un proceso de enseñanza más efectivo.

EDM se ha posicionado como un campo de gran interés para investigadores de diferentes tipos de entornos educativos:

- **Educación tradicional**. En la cual se trata de transmitir conocimientos y habilidades basadas en el contacto personal profesor-alumno y así poder estudiar incluso desde el punto de vista de la Psicología, cómo los alumnos aprenden.

- **El aprendizaje electrónico (*E-learning*)** y los **Sistemas de gestión de aprendizaje (*Learning Management System*, LMS**). El aprendizaje electrónico provee de instrucciones en línea a sus usuarios y los *LMS* proveen de comunicación, colaboración, administración y herramientas para reportar resultados.

- **Sistemas Tutoriales Inteligentes (*Intelligent Tutoring System*, ITS)** y los **Sistemas Hipermedia Adaptativos (*Adaptive Educational Hypermedia System, AEHS*)**. Estos representan una alternativa para colocar en un sitio web, un enfoque educativo que pueda adaptarse a las necesidades de enseñanza de cada alumno.

Figura 2.1 Principales entornos educativos de investigación en EDM.

Hay algunas cuestiones importantes que diferencian la aplicación de la DM al campo educativo, respecto de otros dominios, por ejemplo:

- **El Objetivo.** EDM tiene por objetivos fundamentales mejorar y guiar el proceso de aprendizaje de los estudiantes, además realizar investigación que permita profundizar en el entendimiento de los fenómenos educativos. Estas metas son en ocasiones difíciles de cuantificar y requieren de un propio conjunto de técnicas de medición.

- **La Información.** En los ambientes educativos hay diferentes tipos de información disponible. La información es específica al área educativa y tiene intrínseca información semántica, relaciones con otros tipos de información y múltiples niveles de significado jerárquico.

- **Las Técnicas.** Los problemas educativos y su información tiene algunas características que requieren ser tratados de una manera particular. Aunque muchas de las técnicas tradicionales pueden ser aplicadas directamente, otras no es posible y deben ser adaptadas al problema específico a tratar.

El conocimiento obtenido mediante EDM puede ser usado por diferentes actores del proceso educativo, aunque en consideraciones iniciales se veía que iba dirigido para los estudiantes y los profesores, actualmente hay más interesados y con diferentes objetivos, como los diseñadores de cursos, los investigadores de los procesos educativos, el personal directivo-administrativo de las instituciones, etc. La Figura 2.2 muestra un esquema del proceso general de EDM y a sus actores principales.

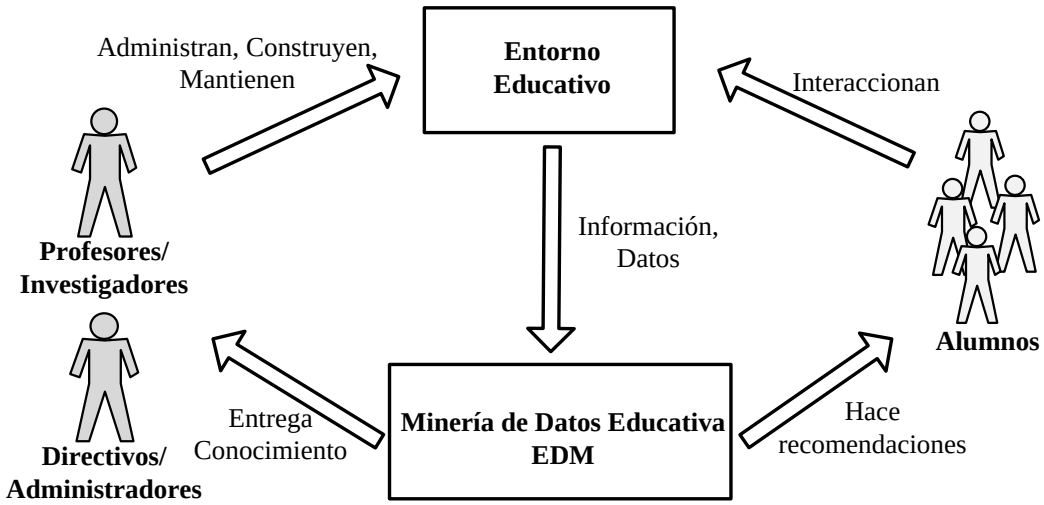


Figura 2.2 Esquema del Proceso de aplicación de EDM.

Hay muchos campos o tareas en el ámbito educativo que pueden ser tratados usando la DM, según Baker (Baker & Yacef, 2009) y Castro (Castro, Vellido, Nebot, & Mugica, 2007) entre las principales están: estudiar sobre el soporte pedagógico que aporta el software de aprendizaje, la investigación científica sobre el aprendizaje y los alumnos, las aplicaciones para la evaluación del desempeño del aprendizaje de los alumnos, las aplicaciones que promueven la adaptación al curso y recomendaciones para el aprendizaje basadas en el comportamiento de los alumnos, los métodos para evaluar el material de aprendizaje y los cursos basados en la web, las aplicaciones para que se dé la retroalimentación alumno-profesor en los cursos en línea y las aplicaciones para detectar a alumnos con comportamientos atípicos. En este mismo sentido, Romero (Romero & Ventura, 2010) profundiza un poco más y establece las siguientes categorías:

- **Análisis y visualización de la información**. El objetivo es resaltar la información útil que sirva como soporte para la toma de decisiones. La

Estadística y la visualización de la información son las técnicas que principalmente son más usadas en este campo.

- **Proveer de información para la retroalimentación como apoyo a los profesores**. Tiene por objetivo retroalimentar a los profesores y administradores con información que les permita tomar decisiones para mejorar el aprendizaje de los alumnos y para organizar los recursos de apoyo a los instructores y profesores de manera más eficiente, también permite que se tomen algunas acciones remediales. Diferentes técnicas de DM han sido usadas en este campo, aunque las reglas de asociación son las más comúnmente utilizadas.

- **Recomendaciones para los estudiantes**. El objetivo es tener la capacidad de realizar recomendaciones directas a los estudiantes respecto de sus actividades personales, de los enlaces de visitas a sitios web, tareas o actividades académicas que tendrán que realizar, etc., además, tener la disponibilidad de adaptar los contenidos de los cursos, interfaces y secuencias para cada estudiante. Diferentes técnicas de DM han sido utilizadas en este campo, pero las más comunes son las reglas de asociación, el agrupamiento y la secuencia de patrones.

- **Adaptación de modelos educativos para los estudiantes**. El objetivo es desarrollar modelos cognitivos para los estudiantes, considerando sus habilidades y conocimientos. La DM ha sido aplicada para considerar automáticamente las características de los alumnos/usuarios tales como la motivación, la satisfacción, el estilo de aprendizaje, el estado afectivo, etc. Las principales técnicas de DM aplicadas a este campo son las redes Bayesianas.

- **Detección de comportamientos indeseables en los alumnos**. El objetivo es detectar a los estudiantes con algún tipo de problema o con un comportamiento inusual, por ejemplo: realizar acciones erróneas, tener poca motivación, jugar juegos, cometer abusos y/o engaños, abandonar, fracasar académicamente, etc. Diferentes técnicas de DM (principalmente la

Clasificación y el Agrupamiento) han sido aplicadas para detectar a estos alumnos a tiempo y así poder ayudarlos apropiadamente.

- **Agrupamiento de estudiantes**. El objetivo es crear grupos de estudiantes de acuerdo a sus intereses y características personales. A partir de los grupos creados, el profesor/instructor puede construir un particular sistema de aprendizaje acorde a cada grupo y, así, propiciar que el aprendizaje sea más efectivo y que cuente con contenidos que se adapten a cada grupo. Las técnicas de DM usadas en este campo son la Clasificación y el Agrupamiento.

- **Análisis de redes sociales**. Tiene el objetivo de estudiar las relaciones entre diferentes individuos, en lugar de estudiar las relaciones entre atributos individuales o propiedades. En una red social se integran individuos con características e intereses en común, lo cual puede ser de mucha utilidad. Diferentes técnicas de DM se han usado para investigar las redes sociales en el entorno educativo, pero el Filtrado Colaborativo es la más común.

- **Desarrollo de mapas conceptuales**. El objetivo es ayudar a los profesores/instructores a construir mapas conceptuales, los cuales son esquemas gráficos que muestran las relaciones entre conceptos y expresan una estructura jerárquica de conocimiento. Algunas técnicas de DM han sido usadas en este campo pero las principales son las Reglas de Asociación y la Minería de Textos.

- **Construcción de cursos para la web**. El objetivo es ayudar a los profesores/instructores y desarrolladores de cursos basados en la web para que los alumnos puedan acceder a contenidos académicos para su aprendizaje de forma automática. Algunas de las técnicas que se han usado son la Clasificación, el Agrupamiento y los algoritmos de Naive.

- **Planeación y programación de cursos**. El objetivo es mejorar los procesos de la educación tradicional con la planeación de cursos futuros, apoyando a los alumnos con la programación de sus cursos, programando la asignación de recursos, apoyando los procesos de admisión, inscripción y asesorías,

desarrollando el currículo, etc. Diferentes técnicas han sido usadas en este campo, pero las principales son las Reglas de Asociación.

- **Predicción del rendimiento académico de los estudiantes**. El objetivo de la predicción es estimar un valor desconocido de una variable que describe al alumno. En la educación estos valores pueden ser las calificaciones o notas, el conocimiento adquirido o el desempeño académico y estas variables pueden ser numéricas/continuas o categóricas/discretas. La predicción del rendimiento de los estudiantes es una de las más antiguas y populares aplicaciones de la DM en la educación, para lo cual se han utilizado diferentes técnicas, como las redes neuronales, las redes bayesianas, los sistemas basados en reglas, los análisis de regresión y correlación. A continuación se describen brevemente y a manera de ejemplo, algunos trabajos que han utilizado diferentes técnicas de DM para predecir el rendimiento académico de los estudiantes. Aguilar (Aguilar, Chawla, Brockman, Ambrose, & Goodrich, 2014) investigó la viabilidad de usar la información del portafolio electrónico de los estudiantes para predecir el nivel de retención en el primer semestre de ingeniería. Demuestra que aunque el conjunto de datos usado no cuenta con rasgos que indican el compromiso académico de los estudiantes, se pueden obtener resultados razonables para identificar a aquellos que abandonarán. Utiliza distintos algoritmos de clasificación y trabajan con un conjunto de datos desbalanceado. Esta metodología puede ser utilizada para establecer un sistema de alerta temprana y así atender de manera oportuna a los estudiantes en riesgo. Rogers (Rogers, Colvin, & Chiera, 2014) compara directamente una técnica de tabulación que denomina "el método del índice", con la técnica de regresión lineal múltiple para identificar a los estudiantes en riesgo. Se comprueba que el método desarrollado obtiene resultados comparables en términos de exactitud con la regresión lineal. Sugiere que en los ambientes de aprendizaje el método desarrollado es una herramienta prometedora para los profesores-educadores, quienes requieren un algoritmo flexible y adaptable. Jiménez (Jiménez, Luna, & Ventura, 2013) trata de predecir lo más temprano posible a los estudiantes que fracasaran en un centro escolar de educación secundaria en España. Para lo anterior se utilizan varios algoritmos de clasificación, lo anterior para poder apoyar lo más temprano posible a los estudiantes en riesgo, con la

finalidad de evitar que fracasen. Kotsiantis (Kotsiantis S. B., 2012), realiza la predicción de las calificaciones finales de 354 estudiantes de la "*Hellenic Open University*" en la materia de Introducción a la Informática, utilizando un modelo de regresión y 17 atributos con información personal de los estudiantes, del tutor de clase y de una prueba final presentada en el aula. Sus mejores modelos obtienen 12% de error. Quadril (Quadril & Kalyankar, 2010) analiza el problema de predecir el abandono escolar de los estudiantes utilizando un modelo híbrido con árboles de decisión y regresión logística. Hace uso de información personal de los estudiantes, de notas previas e incluso de sus padres y encuentra cuáles son las variables que más influyen para que el estudiante decida abandonar la escuela.

2.2.1 UTILIZACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR EL ABANDONO ESCOLAR

Predecir a los estudiantes que fracasarán en sus estudios, ya sea por reprobar o por abandonar, es un problema muy importante a resolver para los profesionales de la educación, además de que al tratarlo ayuda a entender mejor por qué los alumnos fracasan y no completan sus estudios. Como se ha visto en las secciones anteriores una de las dificultades principales de este problema radica en la gran cantidad de factores o características de los estudiantes que pueden afectar su rendimiento académico, como son sus antecedentes académicos, el entorno familiar, sus actividades sociales, su estatus socio-económico, su perfil psicológico y sus rasgos demográficos-culturales (Aloise-Young & Chávez, 2002). En las décadas pasadas se realizaron una gran cantidad de investigaciones para identificar los principales factores que afectan el rendimiento académico de los estudiantes y que los lleva a fracasar o abandonar en los diferentes niveles educativos: en la educación básica (primaria), la educación secundaria o media y la educación superior o universitaria (Araque, Roldán, & Salguero, 2009). Uno de los trabajos teóricos más reconocidos que explica el problema del fracaso escolar de los estudiantes, sus causas y soluciones es el análisis de Tinto (Tinto, 1987). Este modelo sugiere que el nivel de integración social y académica del estudiante en la institución educativa es el factor más determinante para completar los estudios e identifica algunos aspectos que influyen en la mencionada integración, como el respaldo familiar, las características

personales, las escuelas previas, el rendimiento académico previo y las interacciones entre el estudiante y la institución educativa. Más recientemente, ha aparecido un consenso en cuanto a que la detección y prevención del fracaso escolar de los estudiantes y una temprana intervención hacen mucho más sencillo tratar de remediar el problema (Slavin, Karweit, & Wasik, 1994). Siguiendo esta misma línea, una manera efectiva para detectar el fracaso escolar de los estudiantes es usando técnicas de DM, las cuales ya han sido aplicadas exitosamente en otras áreas, como por ejemplo en el comercio electrónico, donde su uso se ha hecho muy popular. DM es una etapa del proceso más global de Extracción del Conocimiento o Knowledge Discovery (KD), que busca extraer conocimiento previamente desconocido, no trivial, potencialmente útil, válido y comprensible, a partir de grandes cantidades de información (Klösgen & Zytkow, 2002).

En la actualidad existen gran cantidad de ejemplos de aplicación de técnicas de DM para crear modelos de predicción del fracaso escolar y del abandono de los estudiantes (Kotsiantis, Patriarchas, & Xenos, 2010). Técnicas estadísticas, como el análisis de regresión y correlación y algunas técnicas de DM, como la clasificación usando árboles de decisión, las redes bayesianas, las redes neuronales, los algoritmos basados en el k-vecino más cercano y otras, se han usado, evaluado y comparado, para predecir el rendimiento académico de los estudiantes (Hämäläinen & Vinni, 2011). A continuación se describen algunos ejemplos representativos de trabajos que usan las técnicas citadas, agrupadas en dos tipos:

- **Técnicas estadísticas**. El análisis de regresión y correlación se utilizaron para predecir el rendimiento académico de estudiantes que llevan sus cursos en línea, (Wang & Newlin, 2002), para identificar los factores que mejor pueden predecir si los alumnos persistirán en sus cursos en la modalidad de educación a distancia (Parker, 1999) y para predecir la probabilidad de cuáles estudiantes de educación media o preuniversitaria concluyan exitosamente la universidad (McDonald, 2004). El análisis de funciones discriminantes se utilizó para determinar cuáles son los atributos que mejor predicen si los estudiantes terminarán exitosamente sus cursos (Martínez, 2001). Modelos de regresión logística se aplicaron para determinar cuál es el riesgo de que los estudiantes abandonen su carrera universitaria (Araque, Roldán, & Salguero, 2009) y para investigar si el tiempo de ocio y aburrimiento es uno de los

factores que puede predecir si estudiantes de educación media abandonaran la escuela (Wegner, Flisher, Lombard, & King, 2008).

- **Técnicas de DM**. El análisis usando árboles de decisión, una de las técnicas más populares de DM, se ha utilizado para predecir las características de los estudiantes que abandonarán la universidad (Quadril & Kalyankar, 2010) y el bachillerato (Veitch, 2004). Algoritmos más avanzados de DM, como las redes neuronales y *Random Forests* junto con árboles de decisión, han sido aplicados para predecir a los estudiantes que culminarán sus estudios exitosamente. Para ello se clasificaron en las siguientes categorías: alumnos en riesgo bajo, en riesgo medio y en riesgo alto de reprobar (Superby, Vandamme, & Meskens, 2006). También los árboles de decisión y *Ramdom Forests* han sido aplicados para identificar los factores asociados con la persistencia de los estudiantes en carreras universitarias de Ciencias e Ingeniería (Méndez, Buskirk, Lohor, & Haag, 2008). Arboles de decisión, redes neuronales, los algoritmos Naive-Bayes y algoritmos de aprendizaje basados en instancias, se usaron para predecir el abandono en estudiantes universitarios (Kotsiantis & Pintelas, 2005).

Los anteriores trabajos muestran resultados prometedores al considerar aspectos sociológicos, económicos o características educativas, de entre las cuales algunas son las más relevantes para la predicción de los estudiantes con bajo rendimiento académico. Es importante destacar que la mayoría de las investigaciones que utilizan DM para resolver problemas de fracaso escolar, como el abandono y la reprobación, se han aplicado principalmente a la educación superior (Kotsiantis S. B., 2009) y, más específicamente, en los cursos en línea o de educación a distancia (Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009). Sin embargo, existe poca información sobre investigaciones realizadas en educación básica o media; y las que se han encontrado a estos niveles educativos son trabajos en los que utilizan principalmente técnicas o métodos estadísticos, pero no de DM (Parker, 1999), salvo algunas contadas excepciones (Jiménez, Luna, & Ventura, 2013). Es importante también hacer notar que los factores que pueden afectar el rendimiento académico de los estudiantes pueden variar significativamente dependiendo del nivel educativo, es decir, ciertos factores pueden ser determinantes en la educación básica (primaria y secundaria), pero no en la educación superior y viceversa; es necesario investigar ampliamente cuáles son todos estos posibles factores. Por tanto, una vez que se disponga de toda la información disponible es

necesario hacer un proceso de selección de los mejores atributos para identificar y seleccionar los factores más importantes o aquéllos que afectan más el rendimiento académico de los estudiantes (Márquez-Vera, Romero, & Ventura, 2011). Otro problema a tener en cuenta, es que en muchos de los casos, la información que se utiliza para predecir a los estudiantes en riesgo de reprobar o abandonar está desbalanceada, lo que significa que sólo una pequeña parte de los estudiantes fracasan o abandonan y la gran mayoría aprueba y continúa con sus estudios. Algunos métodos propuestos por la comunidad de DM y Aprendizaje Automático o *Machine Learning* (ML) para resolver el problema de la clasificación de conjuntos de datos desbalanceados, son considerar diferentes costos en la clasificación y haciendo un re-muestreo al conjunto de datos original. Concretamente, algunos trabajos muestran que la técnica de considerar diferentes costos en la clasificación es una solución muy efectiva para predecir el abandono de los estudiantes de nuevo ingreso a la universidad (Kotsiantis S. B., 2009), también para mejorar los resultados de la clasificación usando conjuntos de datos de estudiantes universitarios (Dekker, Pechenizkiy, & Vleeshouwers, 2009) y para predecir las calificaciones que obtendrán en el examen final del curso los estudiantes universitarios (Romero, Espejo, Zafra, Romero, & Ventura, 2013).

2.2.2 PREDICCIÓN TEMPRANA DEL PROBLEMA

El abandono y la reprobación afecta a la sociedad, a las familias, a las escuelas y a instituciones de todo el mundo, lo cual provoca graves inconvenientes como son: pérdidas financieras, reducción de los índices de aprobación y la reducción del prestigio escolar (Neild, Balfanz, & Herzog, 2007). Por tanto, es de gran importancia poder predecir lo antes posible a los estudiantes que fracasan en la escuela, para poder tomar medidas paliativas a tiempo y poder ayudar a los alumnos en riesgo. La detección temprana de elementos en riesgo o propensos a abandonar sus cursos es crucial para que las estrategias institucionales de retención sean efectivas. Para tratar de reducir el problema mencionado es necesario detectar los casos de riesgo lo más temprano posible, para poder brindar apoyo con la finalidad de evitar el abandono, pues la intervención temprana facilita la retención de los estudiantes (Heppen & Bowles, 2008).

Seidman desarrolló la fórmula para la retención de los estudiantes (Seidman, 1996):

Retención = Identificación Temprana + Intervención (Temprana + Intensiva + Continua)

La fórmula de Seidman muestra que la identificación temprana de los estudiantes en riesgo y el implementar oportunamente una continua e intensiva intervención son la clave para reducir el nivel de abandono escolar. Entonces, desarrollar y utilizar un Sistema de Alerta Temprana (SIAT) es una buena solución para detectar lo más temprano posible a los estudiantes con alto riesgo de abandonar. Un SIAT es cualquier sistema diseñado para alertar sobre un riesgo potencial, su propósito es prevenir que se presente el problema y pueda ser peligroso (Grasso, 2012). Esta definición se debe a que hay diferentes tipos de SIAT, los cuales se han usado donde la detección es muy importante, por ejemplo en ataques militares, en la prevención de conflictos, en las crisis económicas/bancarias, en ambientes de peligro/desastres, en epidemias tanto humanas como de animales, en los sismos, etc. En el ámbito educativo, un SIAT es un conjunto de procedimientos e instrumentos que se utilizan para detectar de manera temprana las características o atributos que tienen los estudiantes en riesgo de abandonar la escuela, además involucra la implementación de intervenciones apropiadas para hacer que los estudiantes permanezcan en la escuela (Heppen & Bowles, 2008). Según la Subsecretaria de Educación Media Superior de México, un SIAT es un conjunto de procedimientos e instrumentos automatizados que permiten, por una parte, detectar oportunamente a alumnos de educación media superior que están en riesgo de abandonar los estudios y, por otra, poner en marcha con la debida oportunidad, las intervenciones adecuadas para lograr su permanencia en la escuela (SUBSECRETARÍA DE EDUCACIÓN MEDIA SUPERIOR - SEMS, 2013).

Las características o atributos a detectar por un SIAT son aspectos del rendimiento académico de los estudiantes, los cuales pueden reflejar de forma precisa y oportuna el riesgo de abandonar de cada estudiante. Pero detectar estos indicadores es difícil, porque no existe una única razón o motivo por la que los estudiantes abandonen, ya que se sabe que es un problema multifactorial como se ha descrito anteriormente. Por tanto es común observar todos los factores que pueden afectar en el rendimiento académico de los estudiantes hasta que estos han abandonado la escuela.

En años recientes, los esfuerzos por desarrollar y aplicar SIAT en entornos educativos se han incrementado. A continuación se mencionan algunos ejemplos que se han implementado en distintos países:

- La Subsecretaria de Educación Media Superior (SEMS) en México ha establecido diversos programas para apoyar a los jóvenes estudiantes y, entre ellos, ha desarrollado un SIAT basado en la utilización de una plantilla de MS EXCEL (Maldonado-Ulloa, Sancén-Rodríguez, Torres-Valadés, & Murillo-Pazarán, 2011). Este SIAT genera alertas comenzando con tres indicadores: ausentismo, bajo rendimiento y conducta o comportamiento problemático. Para estos indicadores se especifican umbrales críticos, éstos son los niveles para los cuales se considera generalmente una alta probabilidad de abandonar. Una vez que los estudiantes en riesgo son identificados, la escuela los apoya con asesorías académicas, atención sicológica, becas y asesorías para el manejo de conflictos.

- El centro nacional de escuelas secundarias de Estados Unidos de América (*National High School Center*, NHSC) también ha definido una guía y un SIAT (Heppen & Bowles, 2008) basado en una plantilla de Microsoft EXCEL y en dos indicadores: en el rendimiento académico en el curso y en la asistencia. Esta herramienta/archivo puede ser descargada en línea (http://www.betterhighschools.org/docs/EWStool.xls) para introducir la información de todos los estudiantes y entonces el sistema automáticamente calcula y elabora un reporte que muestra el estado de riesgo de los estudiantes (Heppen & Bowles, 2008). A partir de esta herramienta, el Departamento de Educación de Delaware (*Delaware Department of Education*, DDOE) ha implementado un SIAT en los estados de Chicago, Colorado y Texas (Uekawa, Merola, Fernandez, & Porowski, 2010). Usan un modelo multi-variable para determinar cuáles son los indicadores que tienen la mayor correlación con el abandono de los estudiantes. Una vez identificados los valores de corte óptimos de cada atributo o indicador, tanto en la escuela secundaria como en el bachillerato, usan el análisis de la curva ROC (*Receiver Operating Characteristics*).

- Finalmente, tres países de Europa: Austria, Croacia e Inglaterra han desarrollado un SIAT (Vassiliou, 2013). Estos sistemas se basan en un

monitoreo sistemático del ausentismo y de las calificaciones de los estudiantes. En Austria, a los profesores se les solicita que identifiquen a cada estudiante usando un cuestionario y algunas escuelas usan herramientas en línea. En Inglaterra, las autoridades locales están involucradas en desarrollar un programa (*Not in Employment, Education or Training*, NEET) para detectar a jóvenes sin empleo, educación o formación. En Croacia, el SIAT es vinculado con la responsabilidad de la escuela para monitorear el número de clases perdidas por los estudiantes.

Después de revisar estos SIAT, hay que indicar que usar simplemente un archivo de MS EXCEL (Maldonado-Ulloa, Sancén-Rodríguez, Torres-Valadés, & Murillo-Pazarán, 2011) (Heppen & Bowles, 2008) no es lo más apropiado si se tiene una gran cantidad de información disponible de los estudiantes. Para estos casos, técnicas más avanzadas como algunas de DM (además de las estadísticas) pueden ser utilizadas para predecir a los estudiantes que abandonaran la escuela (Uekawa, Merola, Fernandez, & Porowski, 2010) (Vassiliou, 2013). Los modelos estadísticos como la regresión logística y el análisis discriminante fueron las técnicas más usadas en estudios de retención para la identificación de factores y su impacto en el problema del abandono escolar de los estudiantes (Kovacic, 2010). Sin embargo, en los últimos años, EDM ha aparecido como una nueva área relacionada con el desarrollo, investigación y la aplicación de métodos computacionales, para detectar patrones en grandes colecciones de datos educativos, que de otro modo sería muy difícil o casi imposible de analizar debido al enorme volumen de información que existe (Romero & Ventura, 2013). Predecir el rendimiento académico de los estudiantes es una de las más conocidas aplicaciones de EDM, donde el objetivo es estimar cuál será el desempeño académico, el conocimiento o las calificaciones de los estudiantes (Romero & Ventura, 2010) (Romero & Ventura, 2007). La Clasificación es la técnica más comúnmente empleada para resolver este tipo de problema, donde se requiere descubrir un modelo predictivo del desempeño académico de los estudiantes utilizando información de sus antecedentes (Hämäläinen & Vinni, 2011) (Romero, Espejo, Zafra, Romero, & Ventura, 2013). Sin embargo, hacer una predicción temprana del abandono de los estudiantes es una tarea más difícil, debido a que la clasificación tradicional considera que todos los atributos están siempre disponibles y no obtiene buenos resultados por la naturaleza temporal de este tipo de información (Antunes, 2010).

A continuación se citan algunos ejemplos de trabajos que han tratado de predecir lo más temprano posible a los estudiantes en riesgo de fracasar. Kovacic (Kovacic, 2010) aplicó tres métodos de clasificación para la predicción temprana del rendimiento académico de los estudiantes (CHAID, exhaustive CHAID y QUEST y CHAID). Utilizó variables sociodemográficas como la edad, el género, los rasgos étnicos, educación, si tiene empleo y la disponibilidad; además consideró otras variables del entorno, como el programa del curso y el bloque del curso, los cuales pueden influir en los estudiantes de la Universidad Politécnica de Nueva Zelanda. Encontró que el mejor método para detectar los factores más importantes que separan a los estudiantes que tienen éxito y los que no fue el del Árbol de Clasificación y Regresión, (*Classification and Regression Tree*, CART), el cual arrojó los rasgos étnicos, el programa del curso y el bloque del curso. Otro análisis comparativo de diversos métodos de clasificación (redes neuronales, árboles de decisión, máquinas de vectores de soporte y regresión logística) fue usado para desarrollar un modelo para detectar lo más temprano posible a los estudiantes de primer año que tienen mayor probabilidad de abandonar (Delen, 2010). La información para este estudio es de una universidad pública de la región oeste de Estados Unidos de América. Los mejores resultados se obtuvieron por las máquinas de vectores de soporte, seguido por los árboles de decisión, luego por las redes neuronales y finalmente por la regresión logística. En otro trabajo similar (Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009), diferentes técnicas de clasificación (redes neuronales retroalimentadas, máquinas de vectores de soporte, ensamble probabilístico simplificado difuso *«probabilistic ensemble simplified fuzzy»* y esquemas de decisión) fueron aplicados para predecir el abandono en cursos en línea usando información de estudiantes de la universidad de Atenas. La mejor técnica en cuanto a rapidez y precisión fue el esquema de decisión. Por otro lado, la Clasificación considerando diferentes costos, fue usada para la predicción temprana del abandono de los estudiantes en la Universidad de Masaryk (Bayer, Geryk, Obsivac, & Popelinsky, 2012). En este trabajo se enriqueció la información de los estudiantes con datos sobre su comportamiento social, obtenidos del correo electrónico y de foros de discusiones. Utilizaron sociogramas y análisis de redes sociales o *Social Network Analysis* (SNA) para obtener esta nueva información. Concluyen que cuatro semestres es el periodo en el que el modelo obtenido puede predecir el abandono de los estudiantes de manera precisa. El algoritmo CAR (*Class Association Rules*), el cual entrega una regla en la que en el consecuente sólo aparece

una proposición relacionada al atributo de la clase, fue aplicado también para la predicción temprana del abandono escolar de los estudiantes por Antunes (Antunes, 2010). El conjunto de datos utilizado en este estudio proviene de los estudiantes inscritos en los últimos cinco años de un programa del Instituto Superior Técnico de Lisboa. El conjunto de datos usado cuenta con 16 atributos sobre ejercicios realizados semanalmente, pruebas y exámenes. Por otro lado, un sistema para el apoyo en la toma de decisiones o *Decision Support System* (DSS) fue desarrollado para predecir el éxito, la excelencia y la retención de los estudiantes del primer año de un programa de educación terciaria (Mellalieu, 2011). Este sistema estaba basado en reglas y ecuaciones de regresión derivadas de un conjunto de datos de prueba, de resultados previos obtenidos por los estudiantes. Finalmente, otro SIAT fue desarrollando usando un Sistema de Administración del Aprendizaje (*Learning Managment System*, LMS) y datos de seguimiento de un curso de educación superior (Macfadyen & Dawson, 2010). Se identificaron 15 variables que mostraban una correlación significativa con las notas finales de los estudiantes de la Universidad de Columbia Británica en 2008. El modelo de regresión generó el mejor ajuste del modelo de predicción para el curso y un análisis de regresión logística binaria demostró, la capacidad de éste modelo de predicción.

2.5 CONCLUSIONES DEL CAPÍTULO

Después de revisar todos los anteriores trabajos, se observa que no existe un consenso sobre cuál es el mejor método o algoritmo para predecir (de manera temprana) a los estudiantes que abandonaran o reprobaran; mientras que algunos reportan un algoritmo particular como el que mejor resultados obtiene, para otros, es justo lo contrario. Los algoritmos tradicionales de clasificación están diseñados para que se obtenga la máxima precisión en el modelo de predicción, pero esto sólo cuando se aplican a conjuntos de datos balanceados, es decir cuando hay un número similar de instancias/estudiantes de cada clase. Sin embargo, cuando se pretende realizar una predicción del abandono escolar de los estudiantes, los conjuntos de datos están más desbalanceados, debido a que normalmente la mayoría de los estudiantes continúan en los cursos y sólo algunos abandonan. En estas condiciones, la precisión puede ser engañosa, porque los algoritmos de clasificación pueden

obtener una alta precisión general, en la que se clasifica muy bien a la clase mayoritaria y la clase minoritaria es prácticamente ignorada. Por lo tanto, es necesario diseñar estrategias o algoritmos específicos capaces de clasificar adecuadamente a la clase minoritaria, como es el caso de la predicción del abandono escolar, en el cual, el mayor interés es clasificar adecuadamente a los estudiantes que abandonarán. Finalmente, también se ha mostrado que la mayoría de los trabajos de investigación relacionados a este tema se dan en el nivel de educación superior o universitaria y sólo pocos se dan en los niveles de educación básica y media. Por otro lado, el predecir de manera temprana a los estudiantes en riesgo de abandonar la escuela es un problema más específico y nada sencillo de resolver. Por un lado, como ya se ha mencionado anteriormente es multifactorial y, por otro, los métodos y técnicas usadas para resolverlo se aplican a posteriori, es decir, cuando se ha conseguido suficiente información para conseguir una buena clasificación y así poder hacer una buena predicción de los estudiantes en riesgo, lo cual puede ser demasiado tarde para tomar decisiones en cuanto a cómo apoyar a estos alumnos y evitar a tiempo su fracaso.

3. PREDICCIÓN DEL FRACASO ESCOLAR UTILIZANDO DIFERENTES TÉCNICAS DE MINERÍA DE DATOS

En este capítulo, se propone un algoritmo de programación genética y otras técnicas de DM para resolver el problema de predecir a los estudiantes que fracasan a su paso por la escuela. Se utiliza información real, obtenida de estudiantes de México. Primeramente se realiza una selección de los mejores atributos con el fin de reducir el problema de la alta dimensión del conjunto de datos. Para resolver el problema del desbalanceo de la información, se realiza un rebalanceo de datos y una clasificación considerando diferentes costos. Después se usa un algoritmo basado en programación genética y se compara su efectividad con otros algoritmos de tipo "caja blanca", con la finalidad de obtener modelos de clasificación más comprensibles. Finalmente, las salidas de los algoritmos se muestran y comparan para seleccionar aquéllas que con mayor precisión clasifican específicamente a los estudiantes que pueden reprobar o abandonar la escuela.

3.1 MÉTODO UTILIZADO

El método propuesto para predecir a los estudiantes que reprobarán o abandonarán la escuela es muy similar al proceso general de Extracción del Conocimiento, el cual se muestra en la Figura 3.1. Las principales etapas de este método son:

- **Recopilación de la información.** Esta etapa consiste en recoger toda la información disponible de los estudiantes. Para ello, deben detectarse cuáles son todos los factores que podrían afectar en el rendimiento académico de los estudiantes y recopilar la información de las posibles fuentes disponibles. Finalmente toda la información debe ser integrada en un conjunto de datos.

- **Pre-procesado.** En esta etapa, el conjunto de datos se prepara para la posterior aplicación de las distintas técnicas de DM. Se aplican tareas típicas de pre-procesado como la limpieza de datos, la transformación de variables y obtener particiones del conjunto de datos. Otras técnicas algo más específicas como la selección de atributos y el rebalanceo de datos pueden ser aplicadas para resolver los problemas de alta dimensión y desbalanceo que suelen presentarse en este tipo de conjunto de datos.

- **Minería de Datos.** En esta etapa, ya con el conjunto de datos pre-procesado, son aplicadas las diferentes técnicas de DM para predecir el fracaso escolar de los estudiantes. Para este problema se utiliza la clasificación de entre las diferentes técnicas que existen. Se propone un algoritmo de clasificación basado en programación genética y se comparan sus resultados con otros algoritmos de clasificación clásicos que descubren modelos basados en reglas de clasificación y en árboles de decisión. Además, se aplica la clasificación sensible a costes que considera diferentes costos por error de cada clase, con el fin de resolver el problema del desbalanceo del conjunto de datos.

- **Interpretación.** En esta última etapa son analizados los modelos de clasificación descubiertos en la etapa anterior. Se analizan las salidas de los distintos algoritmos y se verifican cuáles son los factores que aparecen en las reglas y en los árboles de decisión y cómo se relacionan. A partir de estos análisis se puede realizar una interpretación del problema y su magnitud, para

la futura toma de decisiones que puedan reducir el problema del fracaso y abandono.

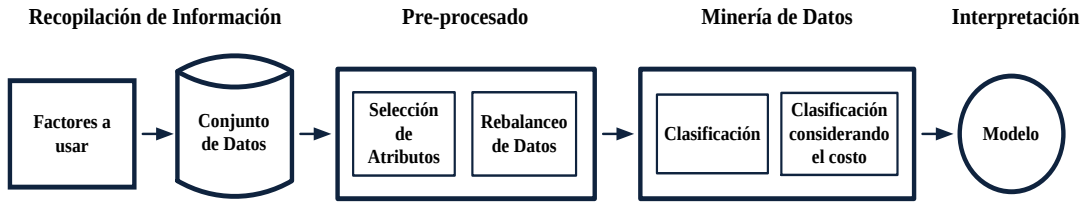


Figura 3.1 Método propuesto para predecir el fracaso escolar de los estudiantes.

3.2 EL CONJUNTO DE DATOS

En esta sección se explica detalladamente cuál es la información de los estudiantes que fue recogida e integrada al conjunto de datos de trabajo. De la misma manera, se explican todas las tareas que fueron realizadas para preparar o pre-procesar los datos, para posteriormente pasar a la etapa de la aplicación de las técnicas de DM.

3.2.1 INFORMACIÓN DE LOS ESTUDIANTES

La información utilizada se recopiló de 670 estudiantes inscritos en el Programa II de la UAPUAZ en el año escolar 2009/10. En el sistema educativo mexicano corresponde al nivel medio-superior, en el cual la mayoría de los estudiantes tienen entre 15 y 18 años de edad y son regulares. Esto es debido a que los programas de educación media-superior tienen una duración de 3 años y su finalidad es que se tenga una adecuada formación académica, conocimientos y competencias generales, que les permita seguir estudiando e integrarse a la universidad. Particularmente, la información usada es solamente de alumnos del primer semestre, en el cual, la mayoría tienen 15 ó 16 años de edad y es donde se presenta el mayor porcentaje de deserción/abandono y reprobación. En la UAPUAZ, para poder promoverse al siguiente periodo escolar, es necesario aprobar todas las asignaturas, o bien, se puede reprobar o suspender sólo una de ellas y llevarla en un curso de regularización. En el

primer semestre se deben cursar 7 asignaturas y se utiliza un sistema decimal en las notas (0…10), donde la nota mínima para aprobar es un 6.

Toda la información fue obtenida entre agosto y diciembre del 2010 de tres fuentes diferentes.

- La primera fuente es un estudio que realiza el Centro Nacional de Evaluación (CENEVAL) a los jóvenes aspirantes a ingresar a la UAPUAZ, previo a la aplicación del Examen Nacional de Ingreso I (EXANI I).
- La segunda fuente es una encuesta diseñada y aplicada a todos los estudiantes a mitad de los cursos, con la finalidad de obtener información personal y familiar para identificar algunos factores que pueden afectar el rendimiento académico.
- La tercera fuente es el Departamento Escolar del plantel, del cual se obtuvieron las calificaciones de todos los estudiantes de las distintas asignaturas.

La Tabla 3.1 contiene las variables/atributos que fueron recogidos de las tres fuentes de datos.

Tabla 3.1 Variables/atributos de la información recopilada de los estudiantes.

Fuente	Variables/Atributos
Estudio del CENEVAL	Edad, Sexo, Secundaria de procedencia, Carácter de escuela, Tipo de escuela, Promedio en la secundaria, Ocupación de la madre, Ocupación del padre, Número hermanos, Limitación para hacer ejercicio, Frecuencia de ejercicio, Tiempo de ejercicio, Nota en Lógica, 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 Civismo, Nota en Ética, Nota en Inglés, Promedio EXANI I.
Encuesta aplicada a los estudiantes	Grupo, Número de estudiantes en el grupo, Turno, Número de amigos, Horas de estudio al día, Método de estudio, Lugar de estudio, Espacio para estudiar, Recursos para estudiar, Hábitos de estudio, Estudia en grupo, Estímulo de los padres para estudiar, Estado civil, Tiene hijos, Religión, Sanción administrativa, Carrera a estudiar, Alguien influye en la elección de la carrera, Personalidad, Discapacidad física, Padece alguna enfermedad, Consumo de alcohol, Fumas, Ingreso familiar, Beca, Trabajas,

	Con quien vives, Nivel educativo madre, Nivel educativo padre, Número de hermanos, Orden entre hermanos, Número de habitantes en tu ciudad, Medio de transporte para ir a la escuela, Distancia a la escuela, Atención en las clases, Te aburres en clase, Interés en las clases, Asignatura difícil, Nivel de motivación, Tomas notas en clases, Método de enseñanza, Te dejan mucha tarea, Calidad de la infraestructura de la escuela, Tutor, Los profesores se ocupan de tu rendimiento académico.
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.

Con toda esta información reunida se dispone de un conjunto de datos formado por 77 atributos o variables que caracterizan a uno de los 670 estudiantes. La Tabla 3.2 muestra el número de variables obtenida de cada fuente de información. El conjunto de datos reunido presenta una alta dimensión, lo cual puede influir en la eficiencia y efectividad de los algoritmos de clasificación (Deegalla, 2006). El problema de la alta dimensión es un fenómeno bien conocido en DM que ocurre cuando el modelo predictivo generado no es muy preciso debido a una abrumadora cantidad de características a elegir entre ellas, por ejemplo, cuando se decide que variable o atributo se usa en un nodo de un árbol de decisión. Aunque algunos algoritmos son poco sensibles a este problema, para otros el costo computacional puede ser prohibitivamente alto. En nuestro caso, se va a tratar de resolver este problema durante los experimentos que se realizarán.

Tabla 3.2 Fuentes de información y atributos de los estudiantes utilizados.

Fuente	Encuesta General	Encuesta Específica	Calificaciones
Tipo de información	Factores Socioeconómicos y calificaciones previas	Factores personales, sociales, familiares y escolares	Calificaciones Actuales
Número de atributos	25	45	7

Finalmente, la variable/atributo de salida o la clase a predecir en nuestro problema es el estatus académico o resultado final del estudiante, el cual tiene dos posibilidades: *APROBÓ* (estudiantes que aprobaron el curso) o *REPROBÓ* (estudiantes que tendrían que repetir el curso). Este atributo fue proporcionado por el Departamento Escolar del Plantel II de la UAPUAZ al final del curso.

3.2.2 PRE-PROCESADO DE LOS DATOS

El pre-procesado permite preparar el conjunto de datos para así poder llevar a cabo correctamente la etapa de clasificación. Se debe destacar que el pre-procesado de los datos es una tarea muy importante, ya que la calidad y confiabilidad de la información disponible, influye directamente en los resultados que posteriormente se obtengan al aplicar los algoritmos de DM. En este apartado se describen las tareas realizadas para preparar el conjunto de datos: la integración de la información, la limpieza de los datos, la transformación de variables, la selección de los mejores atributos, el rebalanceo de la información y obtener las particiones del conjunto de datos.

3.2.2.1 INTEGRACIÓN, LIMPIEZA Y TRANFORMACIÓN

Primeramente, toda la información disponible fue integrada en un único conjunto de datos. Durante este proceso, fue eliminada la información de los estudiantes que no estaba completa al 100%. Es decir, aquellos estudiantes que no completaron el estudio del CENEVAL o que no contestaron la encuesta para detectar los factores que afectan el rendimiento académico, fueron excluidos. También se realizaron modificaciones a los valores de algunos atributos, por ejemplo, las palabras que contenían la letra "Ñ" fueron modificadas, dicha letra fue reemplazada por "N", debido a que el software de DM posteriormente usado no reconoce el mencionado carácter. Se creó un nuevo atributo para la edad en años de cada estudiante, usando la fecha de nacimiento de los mismos. Además, todas las variables de tipo continuo o numérico se transformaron a variables discretas, con el objetivo de proporcionar una visión más comprensible de la información. Por ejemplo, los valores numéricos de las notas obtenidas por los estudiantes en cada asignatura, fueron transformadas de la siguiente manera:

Tabla 3.3 Transformación de las variables del tipo nota.

Excelente	Nota entre 10 y 9.5
Muy bien	Nota entre 9.4 y 8.5
Bien	Nota entre 8.4 y 7.5
Regular	Nota entre 7.4 y 6.5
Suficiente	Nota entre 6.4 y 6.0
Pobre	Nota entre 5.9 y 4.0
Muy pobre	Nota menor a 4.0
NP	No Presentado

Posteriormente, toda la información fue integrada en un fichero de texto con la extensión *.ARFF* de Weka, que es un software de DM (Witten, Frank, & Hall, 2011). A continuación, el conjunto de datos fue dividido de manera aleatoria en 10 pares de ficheros de entrenamiento y de prueba (manteniendo la distribución de clases original). De esta forma cada algoritmo de clasificación podrá ser evaluado usando la técnica de validación cruzada (Kohavi, 1995). Finalmente, tras el pre-procesado de datos se cuenta con un conjunto de datos de 670 estudiantes con 77 variables/atributos y particionado en 10 pares de ficheros.

Este conjunto de datos presenta dos problemas típicos que normalmente se presentan en los conjuntos de datos sobre educación. Por un lado, la alta dimensión, esto es, el número de variables/atributos es muy alto y en estas condiciones usualmente hay algunos que no son muy significativos para la clasificación y es probable que algunas variables/atributos estén relacionadas. Por otro lado, el conjunto de datos está desbalanceado (ver Figura 3.2), esto es porque la mayoría de los estudiantes aprobaron (610) y sólo algunos reprobaron (60).

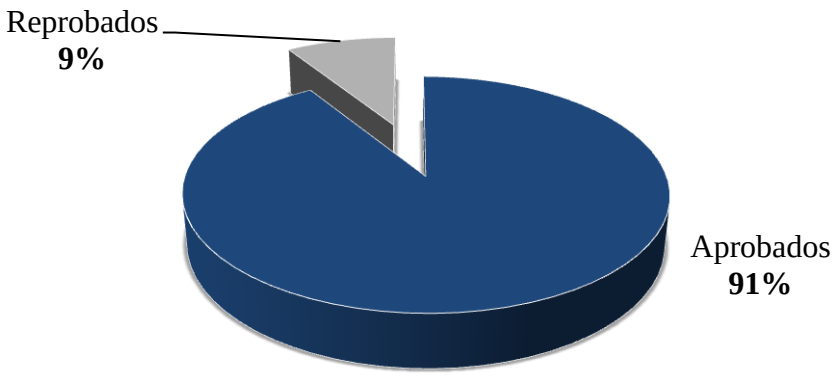


Figura 3.2 Distribución del resultado académico final de los estudiantes.

3.2.2.2 SELECCIÓN DE LOS MEJORES ATRIBUTOS

Se realizó un estudio de selección de mejores atributos para tratar de identificar cuál o cuáles de las características/atributos de los estudiantes tienen un mayor efecto en la variable de salida o clase (Estado Académico). El objetivo es intentar resolver el problema de la alta dimensionalidad del conjunto de datos mediante la reducción del número de atributos sin perder confiabilidad en los resultados de la clasificación. En muchas situaciones prácticas existe una gran cantidad de atributos que los modelos de aprendizaje tienen que considerar y algunos de ellos pueden ser irrelevantes o redundantes. Para ello se usan algoritmos de selección de atributos que tratan de remover aquéllos que son irrelevantes o que influyen poco en la variable de salida. Existe un gran número de algoritmos de selección de atributos que pueden ser agrupados de distintas maneras (Hall & Holmes, 2002).

El software Weka dispone de distintos algoritmos de selección de atributos de entre los que se eligieron diez que se listan a continuación (Witten, Frank, & Hall, 2011): *CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, ReliefFAttributeEval* y *SymmetricalUncertAttributeEval.* Los resultados obtenidos con los datos anteriormente pre-procesados se muestran en la Tabla 3.4. En ella se muestran los mejores atributos seleccionados después de aplicar los citados algoritmos de selección, incluida la frecuencia que representa cuántos de los 10 algoritmos

seleccionaron cada atributo como uno de los que mayor influencia tienen en la variable de salida.

Tabla 3.4 Variables/atributos de mayor influencia organizados por frecuencia de aparición.

Variable/atributo	Frecuencia
Nota en Humanidades 1, Nota en Inglés 1.	10
Nota en Ciencias Sociales 1, Nota en Matemáticas 1, Nota en Taller de lectura y redacción 1, Nota en Física 1, Nota en Computación 1.	9
Nivel de motivación.	5
Promedio en la secundaria	3
Edad, Número de hermanos, Grupo, Fumas, Promedio EXANI I.	2
Estudia en grupo, Estado civil, Tiempo de ejercicio, Nota en Historia.	1

A partir de la tabla anterior fueron seleccionados como mejores atributos aquéllos que tienen una frecuencia mayor a dos, es decir, los que fueron considerados por al menos dos algoritmos de selección de mejores atributos. Con esto se redujo la dimensión del conjunto de datos original, de setenta y siete atributos que se tenían originalmente ahora se cuenta con solamente quince (mejores atributos).

3.2.2.3 REBALANCEO DEL CONJUNTO DE DATOS

El problema de los conjuntos de datos desbalanceados (Gu, Cai, Zhu, & Huang, 2008) surge, debido a que los algoritmos de clasificación tienden a ignorar las clases con menor frecuencia (clases minoritarias) y se enfocan de las clases con mayor frecuencia (clases mayoritarias). Los algoritmos tradicionales de clasificación han sido desarrollados para maximizar la tasa global de precisión, lo cual es independiente de la distribución de clases, esto significa que en la etapa de entrenamiento los algoritmos de clasificación actúan en datos donde la gran mayoría de las instancias corresponden a una clase (clase mayoritaria) y, por ello, cuando se llega a la etapa de prueba se tiene una baja sensibilidad para clasificar a las instancias que corresponden a las clases minoritarias. Una manera para resolver este problema es actuar en la etapa de pre-procesado realizando un rebalanceo (o sobremuestreo) de la distribución de clases. Hay diversos algoritmos de balanceo o rebalanceo de datos, uno de los más usados y que se encuentra disponible en Weka como un filtro de

datos supervisado es el llamado *SMOTE* (*Synthetic Minority Over-sampling Technique*) (Chawla, Bowyer, & Hall, 2002). En el algoritmo *SMOTE* se introducen instancias "sintéticas" de la clase minoritaria entre los segmentos que unen a algunos o a todos los k vecinos más cercanos de la misma clase. Dependiendo del sobre muestreo requerido se introduce aleatoriamente la cantidad de instancias necesarias. En este caso, solamente se re-balancearon los ficheros de entrenamiento (con los 15 mejores atributos) con el algoritmo *SMOTE*, de forma que se introdujeron las instancias necesarias para obtener ficheros de entrenamiento con el 50% estudiantes que aprobaron y 50% de estudiantes que reprobaron, y no se modificaron los ficheros de prueba que no fueron re-balanceados.

Finalmente, después de realizar todas las anteriores tareas de pre-procesado se cuenta con los siguientes ficheros:

- Diez ficheros de entrenamiento y diez ficheros de prueba con todos los atributos (77).
- Diez ficheros de entrenamiento y diez ficheros de prueba con sólo los mejores atributos (15).
- Diez ficheros de entrenamiento y diez ficheros de prueba con sólo los mejores atributos (15); los ficheros de entrenamiento fueron re-balanceados con el algoritmo *SMOTE*.

3.3 MODELOS DE CLASIFICACIÓN Y PROGRAMACIÓN GENÉTICA

En este apartado se describen todos los algoritmos de clasificación usados en la experimentación, incluyendo el algoritmo genético propuesto. Además se explican las diferentes medidas de evaluación que se usaran para cuantificar el rendimiento de los algoritmos.

3.3.1 MODELOS DE CLASIFICACIÓN

Existe un amplio rango de paradigmas que se han utilizado para resolver problemas de clasificación: árboles de decisión, aprendizaje inductivo, aprendizaje basado en instancias y, más recientemente, redes neuronales y algoritmos evolutivos. En este apartado los árboles de decisión y las reglas de inducción son los seleccionados debido a que son técnicas de clasificación de caja blanca, esto es, que proveen una explicación de los resultados de la clasificación y pueden usarse directamente en la toma de decisiones.

- Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica. En un árbol, una instancia se clasifica siguiendo la trayectoria en la que se van cumpliendo condiciones desde la raíz hasta la hoja, la cual se corresponde con una etiqueta de clase.

- Las reglas de inducción generalmente emplean un enfoque que va de lo específico a lo general, en el cual, las reglas obtenidas son generalizadas hasta que se obtiene una descripción satisfactoria de cada clase.

En este capítulo, los 10 siguientes algoritmos de clasificación clásicos, comúnmente usados y disponibles en el software de DM Weka (Witten, Frank, & Hall, 2011) fueron utilizados:

- Cinco algoritmos de reglas de inducción: JRip (Cohen, 1995) el cual implementa un generador de reglas proposicionales, NNge (Roy, 2002) el cual primero clasifica y luego generaliza basándose en el vecino más cercano, OneR (Holte, 1993) el cual usa el atributo de mínimo error para la predicción de una clase, Prism (Cendrowska, 1987) el cual es un algoritmo para la inducción de reglas modulares y Ridor (Richards, 2009) el cual es una implementación del generador de reglas RIpple-Down.

- Cinco algoritmos de árboles de decisión: J48 (Quinlan, 1983) el cual es un algoritmo que genera un árbol de decisión que utiliza la razón de ganancia para seleccionar el atributo de cada nodo, SimpleCart (Breiman, Friedman, Olshen, & Stone, 1984) el cual implementa la mínima poda de costo-complejidad, ADTree (Freund & Mason, 1999) que es un árbol de decisión con una estructura alterna en el que cada nodo sólo hay una bifurcación,

RandomTree (Witten, Frank, & Hall, 2011) el cual considera k atributos elegidos aleatoriamente para cada nodo del árbol y REPTree (Witten, Frank, & Hall, 2011) el cual usa la ganancia de información y realiza una poda de error reducido.

Un árbol de decisión puede ser directamente transformado en un conjunto de reglas del tipo *SI – ENTONCES* (como las que se obtienen de los algoritmos de reglas de inducción), las cuales son las formas más populares de representación de conocimiento debido a su simplicidad y comprensibilidad. De esta manera, un usuario no experto en DM, como un profesor o un directivo escolar, puede usar e interpretar directamente la salida de clasificación de los algoritmos para detectar a los estudiantes con problemas (los de la clase *REPROBÓ*) y poder tomar decisiones de cómo ayudarlos y prevenir su posible fracaso.

Para evaluar el rendimiento de los algoritmos de clasificación, normalmente se utiliza la matriz de confusión. Esta matriz contiene información de las instancias de la clase actual y de la predicción realizada (ver Tabla 3.5).

Tabla 3.5 Matriz de Confusión.

Pred./Act.	**Positivo**	**Negativo**
Positivo	Verdadero Positivo (*True Positive*, **TP**)	Falso Positivo (*False Positive*, **FP**)
Negativo	Falso Negativo (*False Negative*, **FN**)	Verdadero Negativo (*True Negative*, **TN**)

A partir de los resultados de la matriz de confusión se pueden obtener varias medidas de evaluación para cuantificar el rendimiento de los algoritmos de clasificación. En nuestro caso, se utilizaron las siguientes cuatro:

- **Precisión** (*Accuracy*, *Acc*) es la tasa global de precisión de clasificación o precisión de clasificación y se calcula con la siguiente expresión:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$

- **Tasa de verdaderos positivos** (TP_{rate}), también llamada Sensitividad (*Se*), es la razón de las instancias clasificadas como positivas y el total de instancias positivas y se puede calcular con:

$$TP_{rate} = \frac{TP}{TP+FN} \quad (3.2)$$

- **Tasa de verdaderos negativos** (TN_{rate}), también llamada Especificidad (*Sp*), es la razón de las instancias clasificadas como negativas y el total de instancias negativas y se puede calcular con:

$$TN_{rate} = \frac{TN}{TN+FP} \quad (3.3)$$

- **Media Geométrica** (*Geometric Mean*, *GM*), es una medida de tendencia central que usualmente se usa en los conjuntos de datos desbalanceados, indica que tan balanceada es la clasificación de las clases mayoritaria y minoritaria, se puede calcular con:

$$GM = \sqrt{TP_{rate} TN_{rate}} \quad (3.4)$$

Finalmente, además de los 10 algoritmos clásicos de clasificación también se utilizó un algoritmo evolutivo específico llamado "*Interpretable Classification Rule Mining*" (ICRM), el cual se describe en la siguiente sección.

3.3.2 PROGRAMACIÓN GENÉTICA PARA CLASIFICACIÓN

Los algoritmos evolutivos están basados en la teoría de la evolución de Darwin, donde cada individuo codifica una solución y evoluciona para ser un mejor individuo por medio de operadores genéticos (mutación y cruce). La Programación Genética o *Genetic Programming* (GP) es un paradigma de computación evolutiva para encontrar programas computacionales que realicen una tarea definida por el usuario. Se trata de una especialización de los algoritmos genéticos, donde cada individuo es un programa de computadora. Por tanto, puede considerarse como una técnica de aprendizaje automático usada para optimizar una población de programas de ordenador según una heurística definida en función de la capacidad del programa para realizar una determinada tarea computacional definida por el usuario. La programación genética ha sido aplicada exitosamente en problemas complejos de optimización, búsqueda y clasificación (Diosan, Rogozan, & Pecuchet, 2012) (Pan, 2012).

El algoritmo evolutivo propuesto en este capítulo es una variante de un algoritmo genético conocido como Programación Genética Basada en Gramáticas o *Grammar-Based Genetic Programming* (G3P) (Whigham, 1996) en el cual es definida una gramática y el proceso evolutivo garantizando que cada individuo generado cumple dicha gramática. En este caso, se utilizó un algoritmo específico llamado *Interpretable Classification Rule Mining* (ICRM) que emplea el G3P para evolucionar el clasificador generador de reglas y que ha mostrado un buen rendimiento en otros campos de clasificación (Cano, Zafra, & Ventura, 2011). El algoritmo presenta tres variantes que permite conducir el proceso de búsqueda del tipo de reglas en la cuales se está más interesado, por ejemplo, aquéllas que predicen las causas que tienen los estudiantes para reprobar o abandonar sus estudios, lo cual es una capacidad que no tiene ningún otro algoritmo, según la literatura publicada en el tema.

La Figura 3.3, muestra el diagrama de flujo del algoritmo, en el cual, éste realiza iteraciones para encontrar las mejores reglas de las diferentes clases, usando una particular representación de regla. Esta representación provee la mayor eficiencia y aborda el problema de cooperación que trata el orden de las reglas dentro de la base del proceso evolutivo. Se usó una gramática de contexto libre (ver figura 3.4) para especificar qué operadores relacionales son los que pueden aparecer en los "antecedentes" de las reglas y cuales atributos deben aparecer en los "consecuentes" (la clase). El uso de la gramática provee la expresividad, flexibilidad y habilidad para restringir el espacio de búsqueda para encontrar reglas. La implementación de restricciones usando gramática puede ser una manera natural para expresar la sintaxis de las reglas cuando la representación individual ha sido especificada. Los operadores relacionales para atributos nominales son el igual (=) y el diferente (≠). Por tanto, las reglas generadas comprenden una combinación de condiciones de atributos nominales. Estas reglas pueden ser aplicadas a un gran número de problemas como el nuestro. Las reglas son construidas para encontrar la conjunción de condiciones de los atributos más relevantes y que mejor discrimina una clase de las otras. Para ello, hay un operador genético que emplea los atributos que aún no han sido considerados en la regla para encontrar la mejor condición en cualquier otro atributo que cuando se anexa a la regla mejora su precisión. El proceso iterativo continúa mientras que el operador genético no encuentra ningún nuevo atributo o condición relevante que mejore la precisión de la regla o hasta que todos los atributos han sido cubiertos.

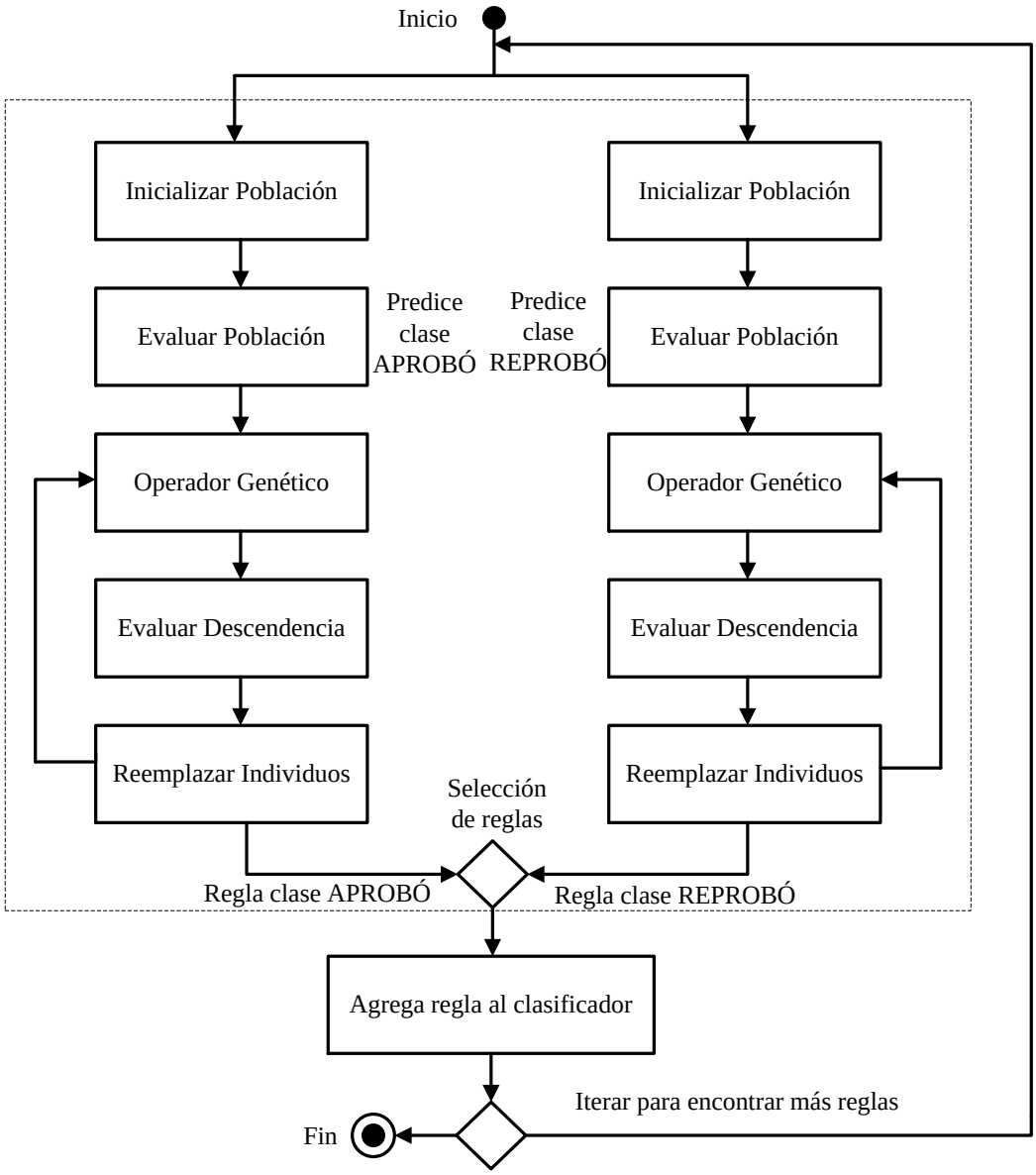


Figura 3.3 Diagrama de flujo del algoritmo ICRM.

<regla>	::=	"**SI**" <antecedente> "**ENTONCES**" <consecuente>
<antecedente>	::=	<antecedente> "**Y**" <condición> | <condición>
<consecuente>	::=	<clase> "**DE OTRO MODO**" <clase> | <clase> "**DE OTRO MODO**" <regla>
<condición>	::=	<atributo> <operador> <valor>
<atributo>	::=	Cualquier atributo válido.
<operador>	::=	"=" | "!="
<valor>	::=	Cualquier valor válido para el atributo correspondiente.
<clase>	::=	"CLASE" = <valorclase>
<valorclase>	::=	"APROBÓ" | "REPROBÓ"

Figura 3.4 Gramática para la generación de reglas del algoritmo ICRM.

La función de adecuación, ajuste o *fitness function* permite evaluar la calidad de las soluciones o individuos mediante una combinación de dos medidas muy comunes en la clasificación, la Sensitividad (*Se*) y la Especificidad (*Sp*). Estas medidas son calculadas a partir de los resultados de la matriz de confusión. El valor de la función de adecuación es calculada como el producto de la Sensitividad y la Especificidad para maximizar la precisión. Esta función tiene un buen desempeño sin importar si el conjunto de datos esta balanceado o desbalanceado.

$$Fitness = Se \cdot Sp \tag{3.5}$$

Finalmente, existen varias maneras de construir la base de reglas. En este capítulo se proponen tres diferentes versiones del algoritmo ICRM para obtener la base de reglas y que sean precisas y útiles para el usuario que busca información de las causas por las que los estudiantes reprueban o desertan.

- La primera versión (ICRM v1) establece una regla por clase. El algoritmo es capaz de obtener reglas precisas para ambas clases, pero finalmente decide qué regla es utilizada en primer lugar (la más precisa) y la otra clase se predice por defecto.
- La segunda versión (ICRM v2) establece varias reglas para cada clase. El algoritmo funciona de manera similar que la primera versión, pero cuando las reglas de ambas clases son obtenidas coloca la mejor regla en el clasificador y elimina del conjunto de datos las instancias consideradas por la regla. En la siguiente iteración, considera solamente las instancias remanentes. Este proceso continua hasta que todas las instancias del conjunto de entrenamiento han sido cubiertas. La salida de clasificación se compone de varias reglas que tienen diferente consecuente.

- La tercera versión (ICRM v3) se extiende de la segunda, pero se enfoca más en la clase de fracaso de los estudiantes. El mayor interés es obtener reglas de clasificación precisas, pero específicamente aquellas que están relacionadas con los estudiantes que fracasan. Por tanto, se establecen múltiples reglas, donde se construyen para predecir los casos de fracaso, que son los de mayor interés. Por último, si ninguna de las reglas que predicen el fracaso cubren una instancia, la otra clase predice a los estudiantes que aprobaran el curso. El número de reglas requerido para predecir los casos de fracaso es decidido por el algoritmo debido a su capacidad de obtener clasificaciones precisas.

3.4 EXPERIMENTOS

Se llevaron a cabo varios experimentos para probar y comparar las tres versiones del algoritmo ICRM con los diez algoritmos de clasificación clásicos mencionados anteriormente. Para tratar de obtener la mejor precisión de clasificación se han usado diferentes métodos de DM para predecir el fracaso escolar de los estudiantes con un conjunto de datos desbalanceado y de alta dimensión.

3.4.1 EXPERIMENTO 1

En el primer experimento, todos los algoritmos de clasificación fueron ejecutados usando la validación cruzada (10 ficheros o *tenfold cross-validation*) y toda la información disponible, es decir, con el conjunto de datos original que tiene 77 atributos de 670 estudiantes. Los resultados de la clasificación de los algoritmos de este experimento se muestran en la Tabla 3.6. Esta tabla muestra la tasa o porcentaje de clasificaciones correctas para cada una de las dos clases: *APROBÓ* (TP_{rate}) y *REPROBÓ* (TN_{rate}), la precisión global (*Acc*) y la media geométrica (*GM*). Además se muestran el número de reglas (*#Reglas*), el promedio del número de condiciones por regla (*#Condiciones por regla*) y el promedio del número de condiciones del clasificador (*#Condiciones*). Cuanto menor sea el número de reglas y de condiciones, mayor es la simplicidad del clasificador y por tanto más alta la comprensibilidad del modelo descubierto. También se puede ver en la Tabla 3.6 que en general los valores obtenidos son altos para la precisión global, *Acc* (mayores a 85.2%) y TP_{rate}

(mayores a 84.4%) pero bajos para TN_{rate} (mayores a 25.0 %) y *GM* (mayores a 49.9%). El mejor algoritmo en términos de TP_{rate} y *Acc* fue ADTree (99.7% y 97.6% respectivamente), los cuales son valores muy altos para predecir a los alumnos de la clase *APROBÓ*. Sin embargo, nuestro mayor interés está en obtener valores altos de TN_{rate}, es decir clasificar adecuadamente a los estudiantes de la clase *REPROBÓ;* donde ICRM v3 obtuvo el mejor resultado (93.3%) y en cuanto a *GM* lo obtuvo ICRM v1 (91.9%). El algoritmo OneR siempre obtiene la clasificación más sencilla (una regla para predecir una clase y la otra por defecto), pero su predicción necesita ser lo más general posible y así obtener la más alta precisión, sin embargo, el valor de TN_{rate} fue bajo (41.7%).

Las versiones del algoritmo ICRM obtienen un TN_{rate} alto y un pequeño número de reglas y condiciones. Por ello, son modelos precisos y comprensibles para predecir el fracaso escolar de los estudiantes y tienen una apropiada relación precisión-interpretabilidad. De este primer experimento, usando todos los atributos disponibles, se ve que todos los modelos de clasificación obtenidos usan pocos atributos del total de disponibles (77).

Tabla 3.6 Resultados de la clasificación usando todos los atributos.

Algoritmo	TP_{rate}	TN_{rate}	***Acc***	***GM***	***#Reglas***	***#Condiciones por regla***	***#Condiciones***
JRip	97.7	78.3	96.0	87.5	8.0	1.5	12.0
NNge	98.5	73.3	96.3	85.0	31.0	76.0	2356.0
OneR	98.9	41.7	93.7	64.2	2.0	0.5	1.0
Prism	99.5	25.0	93.1	49.9	76.0	1.4	110.0
Ridor	96.6	65.0	93.7	79.2	4.0	1.7	7.0
ADTree	**99.7**	76.7	**97.6**	87.4	21.0	1.7	36.0
J48	97.4	53.3	93.4	72.1	31.0	3.1	98.0
RandomTree	95.7	48.3	91.5	68.0	212.0	4.9	1041.0
REPTree	98.0	56.7	94.3	74.5	44.0	1.8	83.0
SimpleCart	97.7	65.0	94.8	79.7	5.0	12.8	64.0
ICRM v1	94.3	90.0	93.9	**91.9**	2.0	1.5	3.1
ICRM v2	97.5	75.0	95.5	85.0	7.6	1.9	14.7
ICRM v3	84.4	**93.3**	85.2	88.5	4.0	1.1	4.5

3.4.2 EXPERIMENTO 2

En el segundo experimento se volvieron a ejecutar todos los algoritmos de clasificación usando nuevamente la validación cruzada, pero ahora con el conjunto de datos reducido (sólo con los 15 mejores atributos). La Tabla 3.7 muestra los resultados de clasificación de este experimento. Comparando los resultados obtenidos en este segundo experimento con los del primero, es decir la Tabla 3.6 y la Tabla 3.7, se puede ver que en general todos los algoritmos mejoraron en dos medidas (TN_{rate} y *GM*). Además, en las otras dos medidas (TP_{rate} y *Acc*) algunos algoritmos cambiaron un poco, unos mejoraron y otros empeoraron, aunque esta variación es poco significativa. De hecho, los valores máximos obtenidos son ahora mejores que los anteriores en dos medidas (TN_{rate} y *GM*). El algoritmo que obtiene estos valores máximos es ICRM v1 (TN_{rate} = 93.3% y *GM* = 92.5%) y junto con el algoritmo OneR obtienen el menor número de reglas (2), pero con un TN_{rate} mucho mayor y un pequeño número de condiciones. Como puede verse en la Tablas 3.6 y 3.7, los valores de TP_{rate} normalmente son más grandes que los de TN_{rate} esto se debe a que el conjunto de datos está desbalanceado.

Tabla 3.7 Resultados de la clasificación usando los mejores atributos.

Algoritmo	***TP_{rate}***	***TN_{rate}***	***Acc***	***GM***	***#Reglas***	***#Condiciones por regla***	***#Condiciones***
JRip	97.0	81.7	95.7	89.0	5.7	1.5	8.7
NNge	98.0	76.7	96.1	86.7	22.2	14.0	310.8
OneR	98.9	41.7	93.7	64.2	2.0	0.8	1.6
Prism	**99.2**	44.2	94.7	66.2	55.6	1.7	93.8
Ridor	95.6	68.3	93.1	80.8	4.0	1.2	5.4
ADTree	**99.2**	78.3	**97.3**	88.1	21.0	3.0	63.0
J48	97.7	55.5	93.9	73.6	19.9	2.1	43.0
RandomTree	98.0	63.3	94.9	78.8	278.6	3.3	912.2
REPTree	97.9	60.0	94.5	76.6	30.0	1.9	68.4
SimpleCart	98.0	65.0	95.1	79.8	6.9	4.1	29.4
ICRM v1	92.0	**93.3**	92.1	**92.5**	2.0	2.4	4.9
ICRM v2	97.2	71.7	94.9	82.8	8.2	2.1	17.9
ICRM v3	75.9	85.0	76.7	79.0	4.0	0.9	3.8

3.4.3 EXPERIMENTO 3

En el tercer experimento, nuevamente fueron ejecutados todos los algoritmos de clasificación usando la validación cruzada, pero ahora con los ficheros de entrenamiento (con sólo los mejores 15 atributos) re-balanceados al 50% con la salida *APROBÓ* y 50% con la salida *REPROBÓ* y los ficheros de prueba sin re-balancear. Los resultados de esta clasificación aparecen en la Tabla 3.8. Si se analiza y compara esta tabla con la anterior (Tabla 3.7), se observa que más de la mitad de los algoritmos usados incrementaron los valores obtenidos y algunos de ellos incluso obtienen un nuevo valor máximo en casi todas las medidas, excepto en *Acc*: Prism (TP_{rate} = 99.8%), ICRM v3 (TN_{rate} = 98.7%), ADTree (*Acc* = 97.2%) e ICRM v2 (*GM* = 97.0%). Destaca también que los mejores resultados en general son obtenidos por los modelos de ICRM, y que tanto Prism como ADTree tienen menor TN_{rate} y GM que todos los modelos de ICRM. ICRM v3 obtiene el mayor valor de TN_{rate} de todos los algoritmos y en el mismo sentido ICRM v2 obtiene el mayor valor de *GM*. Según la Tabla 3.8, entonces ICRM v3 es capaz de predecir el 98.7% de los estudiantes que fracasaron usando solamente cinco reglas y con un promedio de 1.5 condiciones por regla. ICRM v1 obtuvo solamente dos reglas con 1.5 condiciones por regla y es capaz de predecir al 98.0% de los estudiantes que fracasaron.

Tabla 3.8 Resultados de clasificación usando el conjunto de datos re-balanceado.

Algoritmo	TP_{rate}	TN_{rate}	*Acc*	*GM*	*#Reglas*	*#Condiciones por regla*	*#Condiciones*
JRip	97.7	65.0	94.8	78.8	12.3	1.6	18.7
NNge	98.7	78.3	96.9	87.1	24.2	14.0	338.8
OneR	88.8	88.3	88.8	88.3	2.0	1.0	2.0
Prism	**99.8**	37.1	94.7	59.0	67.4	1.9	127.7
Ridor	97.9	70.0	95.4	81.4	9.3	1.9	17.7
ADTree	98.2	86.7	**97.2**	92.1	45.4	2.8	58.0
J48	96.7	75.0	94.8	84.8	19.9	2.8	137.3
RandomTree	96.1	68.3	93.6	79.6	284.1	3.6	1033.4
REPTree	96.5	75.0	94.6	84.6	66.9	2.2	153.9
SimpleCart	96.4	76.7	94.6	85.5	12.3	5.9	78.0
ICRM v1	93.8	98.0	95.9	95.9	2.0	1.5	3.1
ICRM v2	98.0	96.1	97.1	**97.0**	7.9	1.8	14.4
ICRM v3	86.7	**98.7**	92.7	92.5	5.0	1.5	7.6

3.4.4 EXPERIMENTO 4

Otra manera de resolver el problema de clasificación de un conjunto de datos desbalanceado, es considerar diferentes costos en la clasificación (*cost-sensitive classification*) (Elkan, 2001). El tratar de optimizar los resultados de la clasificación sin considerar el costo de los errores, puede llevar a resultados no óptimos, esto se debe a que los grandes costos pueden resultar de una clasificación errónea de la clase minoritaria. De hecho, en nuestro problema particular, el interés mayor se encuentra en clasificar de la mejor manera a la clase minoritaria, aquellos estudiantes que fracasaron y no tanto en aquellos estudiantes que se promovieron aprobando el semestre. Los costos de clasificación se pueden incorporar en cada algoritmo y ser considerados durante la clasificación. En el caso de que existan dos clases, los costos se incorporan en una matriz de 2x2, en la cual los elementos de la diagonal principal representan los dos tipos de clasificaciones correctas (TP y TN) y en la otra diagonal se encuentran los dos tipos de errores (FP y FN).

La Tabla 3.9 muestra la matriz de costos utilizada con los valores por defecto para dos clases, en la cual la diagonal principal tiene ceros, es decir, que para los valores correctos el costo de clasificación es cero y en la otra diagonal aparecen unos, es decir, tener una clasificación errónea tiene un costo de uno.

Tabla 3.9 Matriz de costos con valores por defecto.

Pred./Act.	Positivo	Negativo
Positivo	0	1
Negativo	1	0

Weka permite a cualquier algoritmo de clasificación la incorporación de costos, usando el algoritmo de *metaclassification* denominado *CostSensitiveClassifier*. En este trabajo se incorporaron a la matriz de costos los valores que se muestran en la Tabla 3.10. Se realizaron varias pruebas con diferentes costos y los mejores resultados de clasificación se obtuvieron con la matriz de costos de la Tabla 3.10. Los valores de la matriz usada indican que es cuatro veces más importante para este caso clasificar correctamente a los estudiantes con la salida *REPROBÓ* que a los estudiantes con la salida *APROBÓ*.

Tabla 3.10 Matriz de costos usada.

Pred./Act.	Positivo	Negativo
Positivo	0	1
Negativo	4	0

En el cuarto experimento, se ejecutaron todos los algoritmos de clasificación usando la validación cruzada, pero considerando ahora los diferentes costos de clasificación introduciendo la matriz de costos sobre el conjunto de datos de mejores atributos. La Tabla 3.11 muestra los resultados de este cuarto experimento. Analizando la Tabla 3.11 y comparándola con la Tabla 3.8 se puede ver que algunos algoritmos obtienen mejores valores en algunas medidas, mientras que en otros algoritmos estas medidas disminuyen, por lo anterior no hay una clara mejora en los resultados de clasificación de este cuarto experimento. El algoritmo JRip obtiene el máximo valor de este experimento en cuanto a clasificar a la clase minoritaria en dos medidas (TN_{rate} = 93.3% y *GM* = 94.6%), que son medidas muy importantes para este trabajo.

Tabla 3.11 Resultados de clasificación considerando diferentes costos de clasificación.

Algoritmo	TP_{rate}	TN_{rate}	*Acc*	*GM*	*#Reglas*	*#Condiciones por regla*	*#Condiciones*
JRip	96.2	**93.3**	96.0	**94.6**	8.6	1.5	13.1
NNge	98.2	71.7	95.8	83.0	21.0	14.0	294.0
OneR	96.1	70.0	93.7	80.5	2.0	1.0	2.0
Prism	**99.5**	39.7	94.4	54.0	53.2	1.6	85.4
Ridor	96.9	58.3	93.4	74.0	6.0	1.7	9.9
ADTree	98.1	81.7	**96.6**	89.0	21.0	2.9	61.5
J48	95.7	80.0	94.3	87.1	41.9	2.8	121.3
RandomTree	96.6	68.3	94.0	80.4	251.3	3.4	863.1
REPTree	95.4	65.0	92.7	78.1	41.5	1.6	64.1
SimpleCart	97.2	90.5	**96.6**	93.6	10.2	5.1	56.5
ICRM v1	92.1	91.7	92.1	91.8	2.0	2.5	4.9
ICRM v2	94.4	86.7	93.7	90.3	3.0	3.0	9.1
ICRM v3	94.0	88.3	93.4	90.9	6.0	1.5	8.8

Sin embargo, los mejores resultados de clasificación de todos los experimentos realizados, fueron obtenidos por los modelos de ICRM usando el conjunto de datos de los mejores atributos re-balanceando los ficheros de entrenamiento, es decir en el tercer experimento, en él, se obtuvo como valor máximo de TN_{rate} = 98.7% y de *GM* = 97.0%. Por tanto, en el experimento 3 se obtuvieron los modelos de clasificación más precisos e interesantes para predecir a los estudiantes que fracasan.

Finalmente, se realizó un ranking promedio con los resultados de la clasificación obtenida en los diferentes experimentos de las medidas que se usaron, con el objetivo de poder comparar cuantitativamente los valores de las Tablas 3.6, 3.7, 3.8 y 3.11 (Tabla 3.12). Este ranking ha sido calculado comenzando con cuatro valores (entre 1 y 4) que muestran la posición que ocupa cada medida obtenida en los diferentes experimentos realizados. Así, una calificación promedio más baja en el ranking representa un mejor resultado obtenido por los algoritmos de clasificación en los cuatro experimentos realizados.

Tabla 3.12 Ranking promedio de los resultados de clasificación.

Clasificación	***TP_{rate}***	***TN_{rate}***	***Acc***	***GM***	***#Reglas***	***#Condiciones por regla***	***#Condiciones***
Todos los atributos.	1.92	3.38	2.46	3.46	1.85	2.23	2.38
Mejores atributos.	2.62	2.69	2.46	2.69	1.62	1.92	2.00
Rebalanceo de datos.	2.23	1.54	2.08	1.54	3.00	2.85	3.15
Considerando el costo.	3.00	2.15	2.54	2.23	2.00	2.31	2.23

Como puede verse en la Tabla 3.12, el mejor valor de TP_{rate} se consigue con el conjunto de datos que tiene todos los atributos, pero los mejores valores de TN_{rate}, *Acc*, y *GM* se obtuvieron cuando se utiliza el conjunto de datos con sólo los mejores atributos y re-balanceando los ficheros de entrenamiento. Por otro lado, el menor número de reglas y condiciones se obtiene cuando se utiliza el conjunto de datos con los mejores atributos.

3.5 MODELOS DESCUBIERTOS.

En esta sección, se muestran algunos ejemplos de modelos de clasificación descubiertos por diferentes algoritmos, con la finalidad de poder comparar su interpretabilidad y utilidad para identificar a los estudiantes en riesgo de abandonar o reprobar. Esto es muy importante en nuestro problema ya que los modelos se podrán utilizar posteriormente para tomar decisiones sobre cómo detectar y ayudar a los

estudiantes en riesgo. Las reglas descubiertas muestran los factores relevantes y cómo se relacionan para llevar a los estudiantes a aprobar o reprobar el semestre. Los valores de estos factores se pueden obtener antes del final del periodo escolar, con excepción de las calificaciones obtenidas en las diferentes asignaturas. Aunque en nuestro caso se usaron las calificaciones finales obtenidas de cada asignatura, también se pueden usar las calificaciones parciales que van obteniendo los estudiantes en diferentes momentos del semestre, en el caso de la UAPUAZ, se tienen tres periodos en los que los profesores reportan resultados. En el siguiente capítulo de la tesis trataremos este problema de la predicción temprana.

Algoritmo NNge

- Clase *REPROBÓ* SI: GRUPO en {1B, 1K, 1U, 1M, 1O, 1H} ^ NIVEL_MOTIVACION en {Bajo, Regular, No} ^ DISCAPACIDAD_FISICA en {Si} ^ FUMAS en {Si, Ocasionalmente} ^ EDAD en {Más de 15} ^ EXANI_I en {Pobre, Muy pobre, Regular, No presentó) ^ MATE1 en {Pobre, No presentó} ^ FIS1 en {Regular, Suficiente, Pobre, No presentó} ^ CS1 en {Bien, Suficiente} ^ HUM1 en {No presentó} ^ TLR1 en {Suficiente, Regular, Pobre, No presentó} ^ ING1 en {Suficiente, Pobre, Regular, No presentó} ^ COM1 en {Regular, Suficiente}.
- Clase *REPROBÓ* SI: GRUPO en {1B, 1O, 1C, 1I, 1L, 1X, 1G} ^ NIVEL_MOTIVACION en {Regular, Bajo, No} ^ DISCAPACIDAD_FISICA en {Si} ^ FUMAS en {Si, Ocasionalmente} ^ EDAD en {15.0} ^ EXANI_I en {Pobre, Muy pobre} ^ MATE1 en {Pobre, No presentó} ^ FIS1 en {Pobre} ^ CS1 en {Suficiente, Regular, Pobre} ^ HUM1 en {Pobre, Bien} ^ TLR1 en {Suficiente, Pobre} ^ ING1 en {Suficiente, Pobre, Regular, No presentó}.
- Clase *REPROBÓ* SI: GRUPO en {1U, 1H} ^ NIVEL_MOTIVACION en {Bajo, Regular} ^ DISCAPACIDAD_FISICA en {Si} ^ FUMAS en {Si} ^ EDAD en {15.0} ^ MATE1 en {No presentó, Pobre} ^ FIS1 en {No presentó} ^ CS1 en {No presentó} ^ HUM1 en {No presentó} ^ TLR1 en {Pobre, No presentó} ^ ING1 en {Pobre}.
- Clase *APROBÓ* SI: GRUPO en {1J, 1I, 1S, 1Q} ^ NIVEL_MOTIVACION en {Alto} ^ DISCAPACIDAD_FISICA en {No} ^ FUMAS en {No} ^ EDAD en {15.0} ^ EXANI_I en {Bien, Muy bien} ^ MATE1 en {Bien, Excelente, Muy bien} ^ FIS1 en {Muy bien, Excelente, Bien} ^ CS1 en {Bien, Excelente, Muy bien} ^ HUM1 en {Bien, Excelente, Muy Bien} ^ TLR1 en {Muy bien, Excelente, Bien} ^ ING1 en {Bien, Muy bien, Excelente} ^ COM1 en {Bien, Muy bien, Excelente}.
- Clase *APROBÓ* SI: GRUPO en {1J, 1Y, 1A} ^ NIVEL_MOTIVACION en {Alto} ^ DISCAPACIDAD_FISICA en {No} ^ FUMAS en {No} ^ EDAD en {15.0} ^ EXANI_I en {Muy bien, Excelente, Bien} ^ MATE1 en {Muy bien, Excelente} ^ FIS1 en {Muy bien, Excelente, Bien} ^ CS1 en { Muy bien, Bien} ^ HUM1 en {Muy bien, Excelente, Bien} ^ TLR1 en {Muy bien} ^ ING1 en {Muy bien, Excelente, Bien} ^ COM1 en {Muy bien, Excelente, Bien}.
- Clase *REPROBÓ* SI: GRUPO en {1K, 1U, 1H, 1M} ^ NIVEL_MOTIVACION en {Bajo, Regular}.

Figura 3.5 Algunas reglas de salida descubiertas por el algoritmo NNge usando rebalanceo de datos.

Las Figuras 3.5 y 3.6 muestran algunos ejemplos de reglas obtenidas por dos algoritmos clásicos, uno de cada tipo, uno de reglas de inducción y otro de árboles de decisión. Por un lado, se puede ver (Figuras 3.5) que las reglas del algoritmo NNge son poco comprensibles debido a que son muy largas y tienen un gran número de condiciones en el antecedente de la regla, además utilizan el operador lógico "O" (OR) (lista de valores entre paréntesis y separadas por comas). Por otro lado, se

observa (Figuras 3.6) que la salida del algoritmo J48 es más comprensible, ya que es mucho menos grande y no utiliza el operador "O".

```
Algoritmo J48
----------------------
MATE1 = Suficiente
|       ING1 = Bien: APROBÓ
|       ING1 = Muy bien: APROBÓ
|       ING1 = Suficiente: APROBÓ
|       ING1 = Excelente: APROBÓ
|       ING1 = Pobre
|       |       HUM1 = Pobre: APROBÓ
|       |       HUM1 = Bien: REPROBÓ
|       |       HUM1 = Excelente: APROBÓ
|       |       HUM1 = Muy bien: APROBÓ
|       |       HUM1 = Regular: REPROBÓ
|       |       HUM1 = Suficiente: APROBÓ
|       |       HUM1 = No presentó: REPROBÓ
|       ING1 = Regular: APROBÓ
|       ING1 = No presentó: APROBÓ
MATE1 = Bien: APROBÓ
MATE1 = Pobre
|       FIS1 = Regular
|       |       HUM1 = Pobre: APROBÓ
|       |       HUM1 = Bien: APROBÓ
|       |       HUM1 = Excelente: APROBÓ
|       |       HUM1 = Muy Bien: APROBÓ
|       |       HUM1 = Regular: APROBÓ
|       |       HUM1 = Suficiente: APROBÓ
|       |       HUM1 = No presentó: REPROBÓ
|       FIS1 = Muy bien: APROBÓ
|       FIS1 = Suficiente
|       |       CS1 = Bien: APROBÓ
|       |       CS1 = Suficiente
|       |       |       COM1 = Bien: APROBÓ
```

Figura 3.6 Algunas reglas de salida descubiertas por el algoritmo J48 usando rebalanceo de datos.

Respecto a los atributos que aparecen en las salidas de los algoritmos mencionados (Tabla 3.13), se puede observar que el algoritmo J48 sólo utiliza atributos de tipo nota. También puede verse que el algoritmo NNge no utiliza solamente atributos de tipo nota, también usa otros como Grupo, Nivel de motivación, Discapacidad Física, Fumas, Edad y Calificación Promedio en el EXANI I. En General, la calificación en Matemáticas 1 es el atributo más importante porque aparece en la parte alta del árbol de decisión y en casi todas las reglas, por otro lado, el promedio general obtenido en la escuela secundaria no aparece en estas salidas, como pudiera esperarse.

Tabla 3.13 Descripción de los factores que aparecen en las reglas de clasificación.

Atributo	Descripción
MATE1	Calificación en Matemáticas 1
ING1	Calificación en Inglés 1
FIS1	Calificación en Física 1
HUM1	Calificación en Humanidades 1
CS1	Calificación en Ciencias Sociales 1
TLR1	Calificación en Taller de Lectura y Redacción 1
EXANI I	Calificación en el EXANI I
EDAD	Edad en años
NIVEL_MOTIVACION	Nivel de Motivación
DISCAPACIDAD_FISICA	Discapacidad Física
FUMAS	Si estudiante fuma
GRUPO	Grupo donde toma clases

Finalmente las Figuras 3.7 y 3.8 muestran algunos ejemplos de reglas descubiertas por los modelos del algoritmo ICRM. Estos modelos presentan una apariencia que es similar a la de un árbol de decisión. La precisión en la clasificación y el pequeño número de reglas y condiciones de los modelos de ICRM contrastan notablemente con otros modelos que obtuvieron similar precisión en la clasificación pero tienen un gran número de reglas y condiciones. Por tanto, son modelos muy comprensibles que facilitan la toma de decisiones por parte de profesores o directivos para prevenir el fracaso de los estudiantes en riesgo. De hecho, las reglas de ICRM son las que mejor predicen el fracaso de los estudiantes, obtuvieron un TN_{rate} = 98.7%, cuando se re-balancearon los ficheros de entrenamiento. Respecto a los atributos que aparecen en los modelos, se puede ver que ICRM v1 utiliza solamente atributos tipo nota, pero ICRM v3 utiliza también otros atributos como el Nivel de Motivación o el Grupo al que asiste a las clases.

SI (HUM1 != No presentó Y FIS1 != Pobre Y MATE1 != No presentó Y ING1 != Pobre Y CS1 != Pobre) ENTONCES (Clase = *APROBÓ*) DE OTRO MODO (Clase = *REPROBÓ*).

Figura 3.7 Algunas reglas descubiertas por ICRM v1 usando los mejores atributos.

SI (MATE1 = Muy pobre) ENTONCES (Clase = *REPROBÓ*)
DE OTRO MODO SI (MATE1 = No presentó) ENTONCES (Clase = *REPROBÓ*)
DE OTRO MODO SI (MATE1 = Pobre Y ING1 = Pobre) ENTONCES (Clase = *REPROBÓ*)
DE OTRO MODO SI (HUM1 = No presentó Y NIVEL_MOTIVACIÓN = Bajo Y GRUPO != 1J) ENTONCES (Clase = *REPROBÓ*)
DE OTRO MODO (Clase = *APROBÓ*)

Figura 3.8 Algunas reglas descubiertas por ICRM v3 usando rebalanceo de datos.

3.6 CONCLUSIÓN DEL CAPÍTULO

Como se ha visto en este capítulo, predecir el fracaso escolar de los estudiantes (ya sea que no aprueben o que abandonen la escuela) es una tarea difícil de conseguir, debido a que se trata de un problema multifactorial, en el cual existen factores personales, familiares, sociales, económicos, escolares, etc., que pueden influir en los estudiantes. Además, generalmente la información disponible normalmente está desbalanceada. Para resolver estos problemas se han usado diferentes técnicas y algoritmos de DM para predecir el fracaso escolar de los estudiantes. Se realizaron varios experimentos con datos reales de una escuela de nivel medio superior de la ciudad de Zacatecas, México (Programa II de la UAPUAZ), con la finalidad de predecir el resultado de los estudiantes al final del curso. Se propuso la utilización de un algoritmo genético para obtener un modelo de clasificación preciso y que proporcione reglas de inducción fácilmente comprensibles. Además se ha demostrado que usar técnicas como la selección de mejores atributos, el rebalanceo de datos y considerar diferentes costos de clasificación pueden ser utilizadas exitosamente para mejorar la precisión de la clasificación.

Respecto al trabajo realizado en este capítulo y a los resultados que se obtuvieron en los diferentes experimentos, se destaca que:

- Los algoritmos de clasificación pueden usarse para obtener una muy buena predicción del rendimiento académico de los estudiantes, en particular diferenciar a los estudiantes que aprobarán y los que no lo harán.
- Se ha demostrado la utilidad de la técnica de selección de mejores atributos cuando se tiene un conjunto de datos de alta dimensión, el cual tiene muchos atributos. En este caso se redujo el conjunto de datos de setenta y siete a quince atributos, obteniendo pocas reglas y condiciones sin perder precisión en la clasificación.
- Se han mostrado dos maneras diferentes de tratar el problema de clasificar un conjunto de datos desbalanceado, la primera re-balanceando la información y la segunda considerando diferentes costos en la clasificación. De hecho con el re-balanceo de la información, se mejoraron los resultados de clasificación en TN_{rate}, *Acc* y *GM*, que son las medidas de mayor interés de este trabajo.

- Los modelos del algoritmo ICRM obtuvieron los mejores resultados de TN_{rate} y *GM* cuando se re-balancea la información. Para este caso particular, fueron capaces de obtener la clasificación más precisa y comprensible usando pocas reglas de clasificación en las que aparecen pocas condiciones. Específicamente obtuvo la mejor predicción de los alumnos que fracasaron con un TN_{rate} = 98.7%.

Respecto al conocimiento extraído de los modelos de clasificación obtenidos, se puede destacar que:

- Los algoritmos de clasificación de caja blanca obtienen modelos de clasificación sencillos, capaces de explicar sus predicciones por medio de reglas del tipo *SI – ENTONCES*. En este caso, los algoritmos de reglas de inducción usadas generan directamente reglas del tipo mencionado y las salidas de los árboles de decisión se pueden transformar sencillamente al mismo tipo de reglas. Las reglas del tipo *SI – ENTONCES* son una de las formas más populares de representación de conocimiento debido a su simplicidad y comprensibilidad. Este tipo de reglas son simples y fácilmente comprensibles por cualquier usuario no experto en DM, como un profesor o directivo escolar y se puede directamente tomar decisiones al respecto.
- Respecto a los factores o atributos específicos relacionados con el fracaso escolar de los estudiantes, hay algunos valores específicos que aparecen con mayor frecuencia en los modelos de clasificación obtenidos. Por ejemplo, los valores de las notas que más aparecen en las reglas son los equivalentes a Pobre, Muy Pobre y No Presentó en las asignaturas de Física 1, Humanidades 1, Matemáticas 1 e Inglés 1. Hay que indicar que se han utilizado las calificaciones de los estudiantes, por dos razones principales. La primera es porque no se obtienen buenas clasificaciones si no se consideran las notas de los estudiantes y la segunda, porque en otras investigaciones de este tipo también se han considerado las notas (Fourtin, Marcotte, Potvin, Roger, & Joly, 2006) (Moseley & Mead, 2008).
- Otros factores que aparecen frecuentemente asociados con el fracaso son el tener más de15 años de edad, el tener más de un hermano, el asistir a clases en el turno vespertino y el presentar un bajo nivel de motivación por estudiar.

Para finalizar, en base al trabajo realizado en este capítulo, en el siguiente se amplía mediante una propuesta para implementar un Sistema de Alerta Temprana (*Early Warning System*, EWS). Este sistema está basado también en técnicas de clasificación y tiene la finalidad de detectar lo antes posible a los estudiantes en riesgo de reprobar o abandonar la escuela sin tener que esperar a finalizar el curso. De esta forma se puedan tomar medidas preventivas lo antes posible y así poder reducir el fracaso escolar de los estudiantes.

4. PREDICCIÓN TEMPRANA DEL FRACASO ESCOLAR USANDO MINERÍA DE DATOS

En este capítulo se propone una metodología para detectar tan pronto como sea posible a los estudiantes en riesgo de abandonar o reprobar el curso, con el objetivo de poder establecer las bases de un SIAT basado en DM. Primero, se ha agrupado la información de estudiantes de México tal y como se ha ido capturado en las diferentes etapas o momentos del periodo escolar. Después, para cada etapa se ha realizado una selección de los atributos de los estudiantes que más influencia tienen para abandonar la escuela. Entonces se han aplicado a los datos disponibles en cada etapa varios algoritmos de clasificación clásicos además de un algoritmo genético a los datos disponibles en cada etapa. Finalmente, se han comparado los resultados de clasificación que obtienen y se muestran algunos modelos obtenidos.

4.1 METODOLOGÍA PROPUESTA

La metodología tradicional para predecir el fracaso escolar de los estudiantes utiliza toda la información disponible tras el final del curso. De esta forma puede conseguirse una muy alta precisión en los modelos de clasificación/predicción. Sin embargo, para obtener una detección temprana de los estudiantes en riesgo, se debe realizar una predicción lo más temprano posible usando la información disponible en el mismo momento o etapa en el que se genera a lo largo del periodo escolar. Siguiendo esta idea, la metodología que se propone en este capítulo trata de descubrir en qué fecha o etapa del periodo escolar se dispone de información suficiente para poder realizar una predicción suficientemente buena o confiable y establecer la alerta temprana de los estudiantes en riesgo. De esta forma, varios modelos de predicción pueden obtenerse en distintos momentos a partir de la información recogida en las diferentes etapas del periodo escolar (ver Figura 4.1).

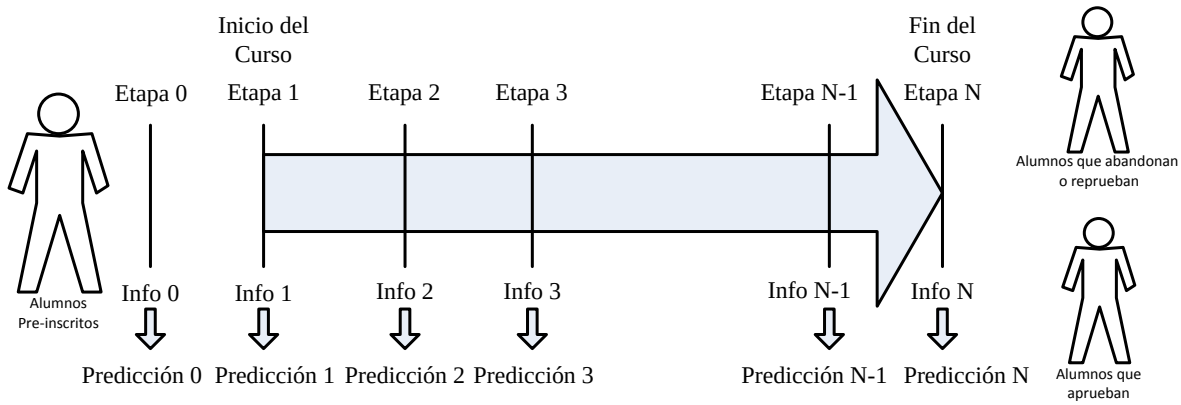


Figura 4.1 Metodología propuesta para predecir de manera temprana a los alumnos en riesgo.

Como puede verse en la Figura 4.1, una vez creado un modelo de clasificación para cada etapa, puede realizarse incluso una predicción incluso antes de iniciarse el curso, para ello se utiliza la información disponible de cursos anteriores y de información personal y administrativa sobre los estudiantes. A medida que el curso va avanzando se dispone de mucha más información de los estudiantes sobre sus actitudes, actividades, resultados académicos, etc. Sin embargo, no es necesario esperar hasta el fin del curso para obtener una buena predicción o lo suficientemente confiable de los estudiantes que abandonan la escuela, los que no aprueban o los que

si lo hacen. El objetivo es encontrar en qué etapa del curso la predicción es lo suficientemente confiable. En la medida que la predicción se haga lo más temprano posible, se puede tener más tiempo para apoyar a los estudiantes en riesgo y por otro lado los profesores y los directivos de la escuela pueden proveer de ayuda específica para corregir a tiempo las actitudes, comportamientos, deficiencias académicas, etc. Para detectar la etapa adecuada, se propone obtener modelos de clasificación en cada etapa (desde la Etapa 0 hasta la etapa N-1) y obtener diferentes medidas de evaluación de la clasificación (como las que se han usado en el capítulo 3) para determinar en cuál etapa se obtiene una clasificación confiable.

4.2 El CONJUNTO DE DATOS

El conjunto de datos utilizado contiene información sobre 419 estudiantes Mexicanos (193 mujeres y 226 hombres) inscritos en el Programa II de la UAPUAZ de primer semestre (Agosto – Diciembre 2012), que es la etapa donde se presenta el mayor porcentaje de reprobación y deserción. En este caso hay un 13.60% de estudiantes que abandonaron (10.3% de mujeres y 20.85% de hombres) como puede verse en la Figura 4.2.

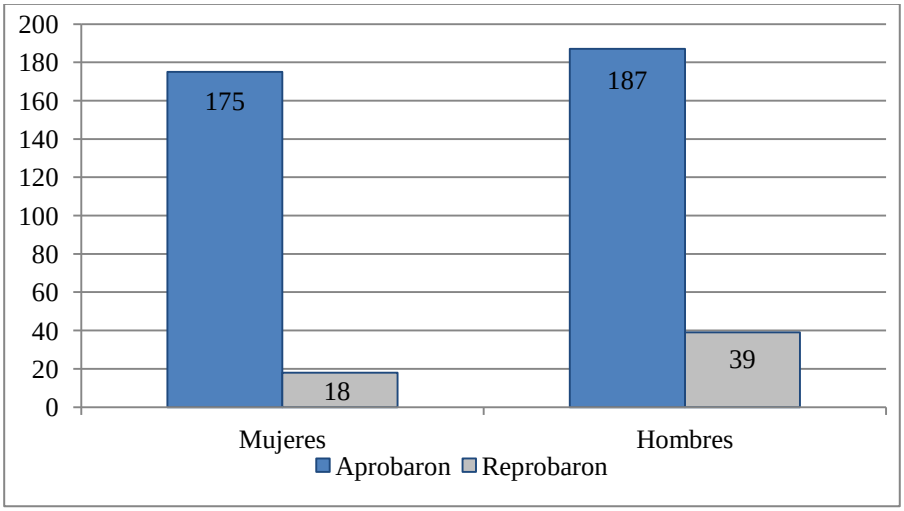


Figura 4.2 Distribución de estudiantes que reprobaron o abandonaron la escuela por género.

Toda la información fue obtenida de distintas fuentes y en diferentes momentos o etapas del semestre entre Agosto y Diciembre de 2012. La Figura 4.3 muestra las etapas específicas que se establecieron y la fechas en las que la información fue recogida.

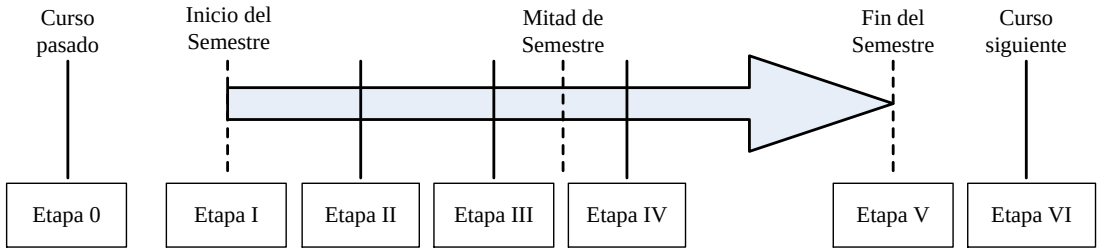


Figura 4.3 Etapas en las que fue recolectada la información.

A continuación, se hace una breve descripción de cuando ocurrieron estas etapas:

- La Etapa 0 fue antes del inicio del semestre y proporciona información de calificaciones previas.
- La Etapa I fue justo al inicio del semestre y contiene información obtenida de la inscripción al Plantel.
- La Etapa II fue cuatro semanas después del inicio del semestre y tiene información sobre las capacidades físicas de los estudiantes.
- La Etapa III fue seis semanas después del inicio del semestre y tiene información sobre asistencia y comportamiento de los estudiantes.
- La Etapa IV fue diez semanas después de iniciado el semestre y tiene una gran cantidad de información sobre factores que pueden afectar el desempeño académico de los estudiantes.
- La Etapa V fue al culminar el semestre y tiene las calificaciones finales obtenidas por los estudiantes de todas las materias del semestre.
- Finalmente la Etapa VI fue justo antes del inicio del próximo semestre y contiene la información sobre qué estudiantes se inscribieron al próximo semestre y qué reprobaron o abandonaron.

La información específica y de la cantidad de atributos usados en cada etapa se muestra en la Tabla 4.1.

Tabla 4.1 Información sobre los estudiantes utilizada en cada etapa.

Etapa	No. de Atributos	Nombre/Descripción
0	2	Promedio General en Secundaria, Calificación en EXANI I.
I	10	Grupo, Número de estudiantes en el grupo, Edad, Turno, Nivel de Ingreso familiar, Beca, Trabaja, Con quien vive, Nivel educativo madre, Nivel educativo padre.
II	11	Discapacidad física, Estatura, Peso, Cintura, Flexibilidad, Abdominales en 1 min, Planchas en 1 min, Tiempo en carrera de 50 m, Tiempo en carrera de 1000 m, Consumo de alcohol, Fumas.
III	4	Asistencia, Aburrimiento en clase, Comportamiento, Sanción Administrativa.
IV	25	Número de amigos, Tiempo adicional dedicado a estudiar, Método de estudio, Lugar de estudio, Hábitos de estudio, Como resuelve sus dudas, Nivel de motivación, Religión, Influencia para elegir carrera universitaria, Personalidad, Recursos para estudiar, Número de hermanos, Orden entre hermanos, Estímulo de los padres para estudiar, Tiempo viviendo en la ciudad, Medio de transporte para ir a la escuela, Distancia a la escuela, Interés en las materias, Asignatura difícil, Tomas notas en clases, Demasiada tarea, Método de enseñanza, Calidad de la infraestructura de la escuela, Tutor, Los profesores se ocupan de tu rendimiento académico.
V	7	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.
VI	1	Reprobó/Continúa en el siguiente semestre.

Como puede verse en la Tabla 4.1, en total se dispone de 60 atributos o indicadores que fueron recolectados en diferentes etapas (desde la Etapa 0 a la V) con el objetivo de predecir qué estudiantes abandonarán o continuarán en la escuela (Etapa VI).

4.3 EXPERIMENTOS

Se realizaron varios experimentos con el objetivo tanto de probar la metodología propuesta como de comparar los resultados del algoritmo propuesto ICRM frente a otros cinco algoritmos de clásicos de clasificación, bien conocidos y disponibles en el software de DM Weka (Witten, Frank, & Hall, 2011), los cuales se describen brevemente a continuación:

- **Clasificadores Bayesianos** (*Bayesian classifier*), Naive-Bayes (John & Langley, 1995). El clasificador Naive-Bayes es un clasificador probabilístico basado en el teorema de Bayes y se caracteriza por sus hipótesis de independencia. En términos simples, este clasificador asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la variable de clase.
- **Máquinas de vectores de soporte** (*Support vector machine*), SMO (Platt, 1998). Implementa el algoritmo de optimización mínima secuencial de Platt para entrenar el clasificador de soporte vectorial usando funciones polinómicas o RBF (*Radial Basis Function*) kernels. Esta implementación globalmente remplaza todos los valores perdidos y reemplaza todos los atributos nominales a binarios. También normaliza todos los atributos por defecto. Los problemas Multi-clase son resueltos usando clasificadores de pares (*pairwise classifiers*).
- **Aprendizaje perezoso basado en instancias**, IBk (Aha & Kibler, 1991). El bien conocido algoritmo del k vecino más cercano se fundamenta en que una nueva instancia se clasifica en la clase más frecuente a la que pertenecen sus k vecinos más cercanos.
- **Reglas de clasificación**, JRip (Cohen, 1995). Implementa el generador de reglas RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*), que fue propuesto por William W. Cohen como una versión optimizada de IREP. Está basado en reglas de asociación con poda para reducir el error, una muy común y efectiva técnica que se utiliza en árboles de decisión.
- **Árboles de decisión**, J48 (Quinlan, 1983). Es la implementación de código abierto del algoritmo C4.5, el cual construye árboles de decisión de un

conjunto de datos de entrenamiento usando el concepto de entropía de la información. En cada nodo del árbol, C4.5 elige el atributo que mejor divide al conjunto de datos en subconjuntos que pertenecen a una clase u otra, sucesivamente divide en subconjuntos más pequeños y toma al atributo con la mayor ganancia de información para tomar la decisión (ocupa un nodo del árbol).

Se propone además la utilización de un algoritmo genético basado en gramática o *Grammar-Based Genetic Programming* (GBGP) para predecir de manera precisa y temprana a los estudiantes que abandonan o reprueban. Este algoritmo es una versión modificada del algoritmo ICRM (Cano, Zafra, & Ventura, 2011) y ha sido adaptada para que obtenga buenos resultados en la clasificación de conjuntos de datos desbalanceados y más específicamente para clasificar-predecir el abandono o reprobación de los estudiantes. El objetivo principal del algoritmo es obtener reglas de clasificación precisas que puedan predecir cuáles son los estudiantes que fracasarán. El algoritmo genera dos conjuntos de reglas del tipo *SI - ENTONCES*, un tipo muestra las condiciones de los estudiantes que aprueban y continúan, y el otro predicen a los que fracasan. Las reglas son obtenidas por medio de un procedimiento evolutivo que construye iterativamente las reglas de clasificación.

Para evaluar el rendimiento de los clasificadores en cada etapa del curso se utilizarán nuevamente las medidas de Precisión (*Acc*), la Tasa de verdaderos positivos (TP_{rate}) o Sensitividad (*Se*), la Tasa de verdaderos negativos (TN_{rate}) o Especificidad (*Sp*) y la media geométrica (*GM*), que ya han sido definidos en el capítulo anterior.

4.3.1 EXPERIMENTO 1

En este primer experimento se usaron todos los atributos en cada etapa del curso, es decir, todos los atributos disponibles desde el inicio del curso en las correspondientes etapas. Los resultados de clasificación obtenidos con todos los algoritmos se muestran en la Figuras 4.4 (TP_{rate}), Figura 4.5 (TN_{rate}), Figura 4.6 (*Acc*) y Figura 4.7 (*GM*).

En la Etapa 0 sólo dos atributos eran conocidos. Esta información fue obtenida antes del inicio del curso y era sobre el rendimiento académico de los estudiantes en cursos previos y del resultado del examen de ingreso. Los valores TN_{rate} y *GM* obtenidos

por el algoritmo ICRM fueron los más altos comparándolos con los otros algoritmos clásicos, siendo la diferencia bastante notable (ver Figuras 4.5 y 4.7). Sin embargo, el algoritmo ICRM obtuvo un valor bajo en la precisión para clasificar a los estudiantes que continuarán el siguiente semestre (TP_{rate}). Por ello, en esta etapa no se recomienda utilizar el algoritmo ICRM para una predicción temprana de los estudiantes que reprueban o desertan. Por otro lado, el resto de los algoritmos obtuvieron un valor muy alto de TP_{rate} cercano a 1.0, pero en su contra obtuvieron un valor muy bajo de TN_{rate} menor de 0.1, es decir, en su predicción prácticamente consideran que todos los estudiantes aprobarán. En base a lo anterior, las predicciones de los algoritmos clásicos en esta etapa no son nada confiables por la gran diferencia de precisión de clasificación en las dos clases de salida.

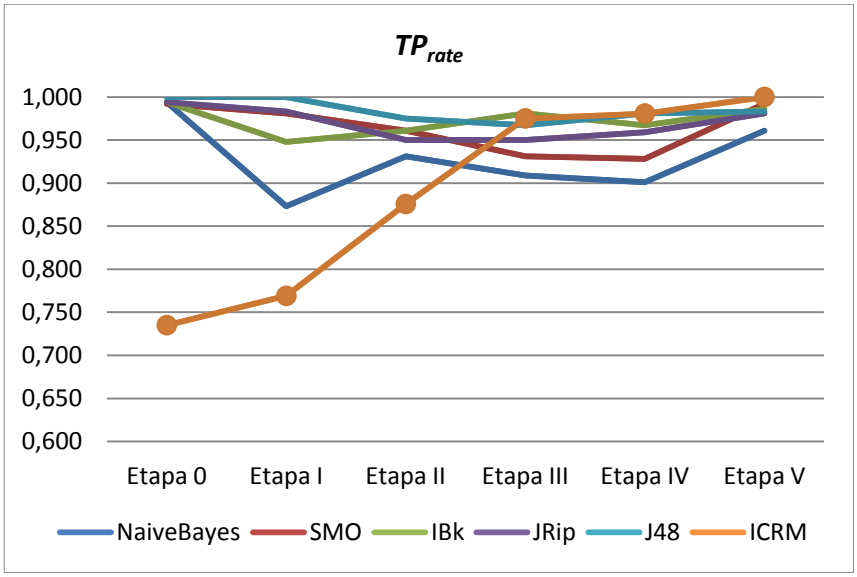


Figura 4.4 TP_{rate} Experimento 1.

En la etapa I, diez nuevos atributos fueron obtenidos y agregados al conjunto de datos anterior para disponer de mayor información sobre las características del grupo donde se encuentran los estudiantes y también sobre algunas características familiares. La nueva información permitió incrementar un poco el valor de TP_{rate} para el algoritmo ICRM y conservó un valor alto de TN_{rate}. Sin embargo, el valor obtenido de TN_{rate} no es todavía suficientemente bueno como para realizar una predicción confiable y poder ser implementada en un sistema de alerta temprana. Por otro lado, algunos de los algoritmos clásicos incrementaron el valor de TN_{rate}, y unos pocos disminuyeron el valor de TP_{rate}.

En la etapa II, once atributos se integraron al conjunto de datos con información sobre las condiciones y capacidades físicas de los estudiantes. Entonces, el algoritmo ICRM redujo significativamente la diferencia con los otros algoritmos respecto al valor obtenido de TP_{rate} (el valor más alto obtenido es mayor de 0.95) y conservó su alto valor de TN_{rate}. El resto de los algoritmos incrementaron significativamente el valor de TN_{rate} (aproximadamente 0.5) y, consecuentemente, el valor de *GM* mejoró a un nivel aceptable (con valores alrededor de 0.7). Así, en esta etapa ya se cuenta con valores que pueden ser confiables con respecto a las medidas de evaluación de la clasificación como para realizar una predicción temprana de los alumnos que pueden reprobar o abandonar, especialmente usando el algoritmo ICRM.

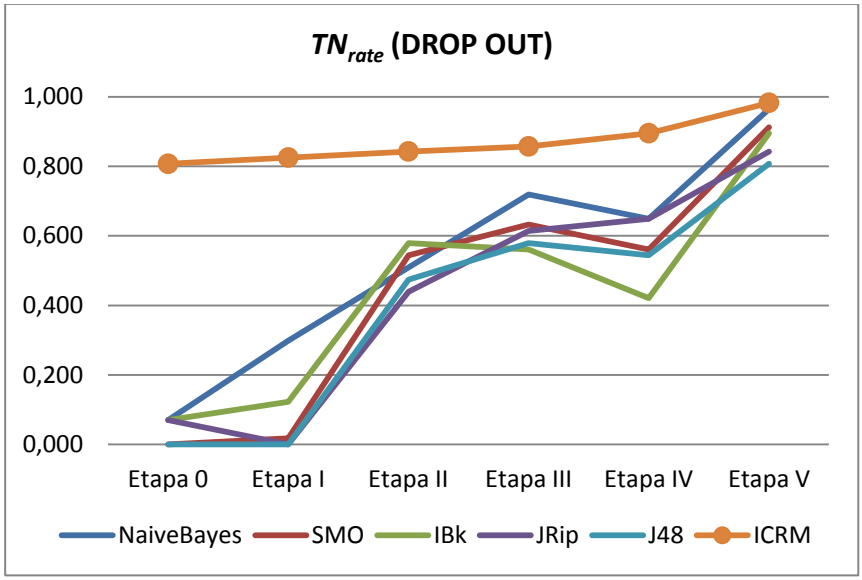


Figura 4.5 TN_{rate} Experimento 1.

En la etapa III se agregaron al conjunto de datos cuatro atributos sobre el comportamiento de los estudiantes en clases. Como puede verse en las Figuras 4.5 y 4.7 los valores de TN_{rate} y *GM* se incrementaron en todos los algoritmos. De hecho, el algoritmo ICRM obtiene un valor muy alto para TP_{rate} (mayor que 0.95) mientras que también conservó un alto valor de TN_{rate} (mayor de 0.8). Este buen desempeño permite recomendar el uso de este algoritmo (por encima de los demás) para realizar una predicción temprana de la reprobación y el abandono de los estudiantes en esta tercera etapa. Hay que indicar, que esta tercera etapa está antes de la mitad del curso, por tanto hay tiempo de margen para tratar de apoyar o ayudar a los estudiantes en riesgo y prevenir que reprueben o abandonen.

En la etapa IV, veinticinco nuevos atributos de información sobre factores que pueden afectar el rendimiento académico de los estudiantes fueron agregados al conjunto de datos. Sin embargo, como puede verse en la Figuras 4.4, Figura 4.5, Figura 4.6 y Figura 4.7 estos nuevos atributos introducen demasiada información que produce ruido en el rendimiento de todos los algoritmos. Por ello, muchos de los algoritmos disminuyeron su eficiencia de clasificación, especialmente en la predicción que más nos interesa, que es el valor de TN_{rate}. Además, esta etapa ocurre después de la mitad del curso, cuando empieza a ser ya un poco tarde para realizar acciones para evitar el fracaso de los estudiantes.

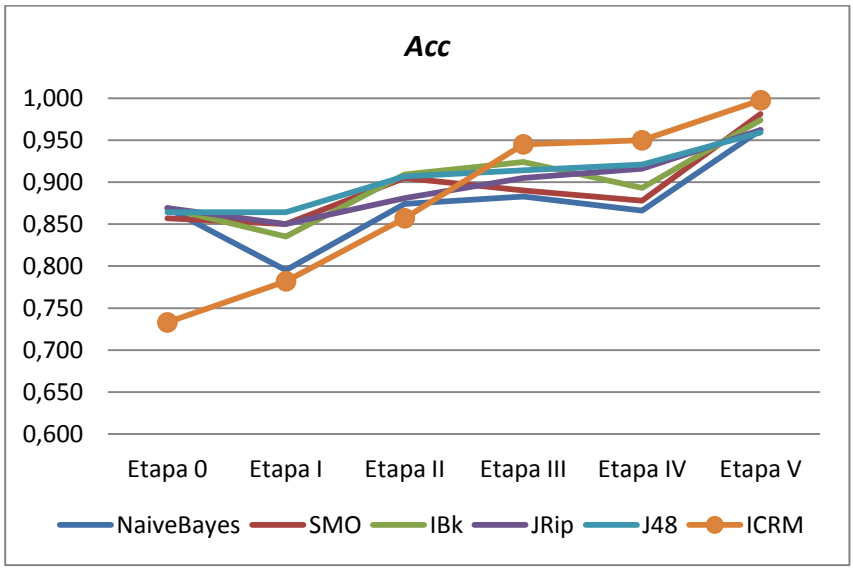


Figura 4.6 *Acc* Experimento 1.

Finalmente, en la etapa V se agrega la información sobre las calificaciones obtenidas en todas las asignaturas del curso. En esta etapa final se puede ver que todos los algoritmos fueron capaces de predecir exitosamente el abandono o reprobación de los estudiantes con valores de TN_{rate} cercanos al máximo. Esto es claramente debido a la alta correlación que existe entre las calificaciones de los exámenes finales y el estado final que tienen los estudiantes. Sin embargo, esta etapa está al final del curso, cuando ya es demasiado tarde para poder intervenir y poder apoyar a los estudiantes que están en alto riesgo.

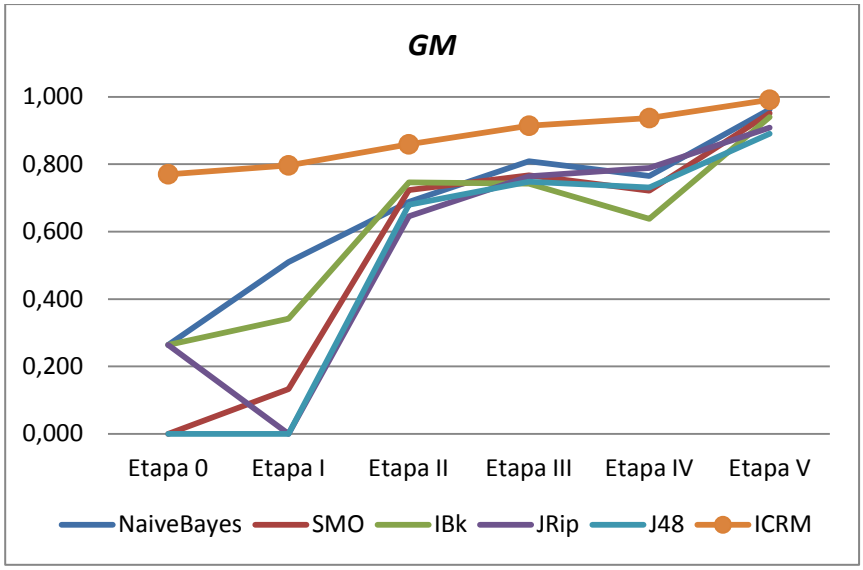


Figura 4.7 *GM* Experimento 1.

4.3.2 EXPERIMENTO 2

En un segundo experimento se llevó a cabo un estudio previo de selección de atributos para identificar cuáles de ellos tienen un mayor efecto para hacer una buena predicción en cada etapa. El objetivo es resolver el problema de la alta dimensión del conjunto de datos sin perder confiabilidad en la clasificación, es decir, encontrar aquellos atributos más significativos o de mayor influencia para nuestra clase de salida o atributo a predecir. Para realizar este selección de los mejores atributos se utilizó el mismo procedimiento descrito en el capítulo anterior, en el cuál se ejecutaron diez algoritmos de selección de atributos proporcionados por el software Weka (Witten, Frank, & Hall, 2011), y posteriormente se realizó un ranking de los resultados obtenidos. Finalmente, se contó el número de veces que cada atributo es seleccionado por cada algo algoritmo y se seleccionó como mejores atributos a aquéllos que fueron seleccionados al menos por dos algoritmos, es decir, aquéllos con frecuencia mayor o igual a dos. La Tabla 4.2 muestra la lista y el número de atributos seleccionados en cada etapa del curso.

Tabla 4.2 Atributos seleccionados como mejores en cada etapa.

Etapa	No. de Atributos	Nombre/Descripción
0	1	Promedio General en Secundaria.
I	6	Grupo, Edad, Turno, Número de compañeros, Trabaja, Nivel educativo madre.
II	3	Tiempo en carrera de 1000 m, Consumo de alcohol, Fumas.
III	2	Asistencia, Sanción Administrativa.
IV	2	Lugar de estudio, Nivel de motivación.
V	3	Nota en Matemáticas 1, Nota en Ciencias Sociales 1, Nota en Humanidades 1.

Cuando se comparan los atributos de las Tablas 4.1 y 4.2 puede observarse que hay una gran reducción en el número de atributos en algunas etapas, como por ejemplo en la Etapa II (donde se pasa de once a tres atributos) y aún más notoria es en la Etapa IV (donde se pasa de veinticinco a dos atributos).

Entonces se ejecutaron nuevamente todos los algoritmos de clasificación de la misma manera que en el primer experimento, pero ahora usando sólo los mejores atributos seleccionados, esto es, sólo con los atributos seleccionados que se van agregando desde la etapa inicio del curso hasta cada etapa. Los resultados de la clasificación obtenida en el segundo experimento se muestran en la las Figura 4.8 (TP_{rate}), Figura 4.9 (TN_{rate}), Figura 4.10 (*Acc*) y Figura 4.11 (*GM*), donde nuevamente aparecen las medidas que ya se han utilizado.

En la etapa inicial sólo el promedio de calificaciones obtenidas en la secundaria fue seleccionado como el atributo más relevante. Cuando se comparan los resultados usando los mejores atributos del segundo experimento y usando todos los atributos del primer experimento, puede verse que los valores de TP_{rate} y *Acc* son similares en todos los algoritmos, mientras que los valores de TN_{rate} y *GM* son menores en casi todos los algoritmos excepto el algoritmo ICRM.

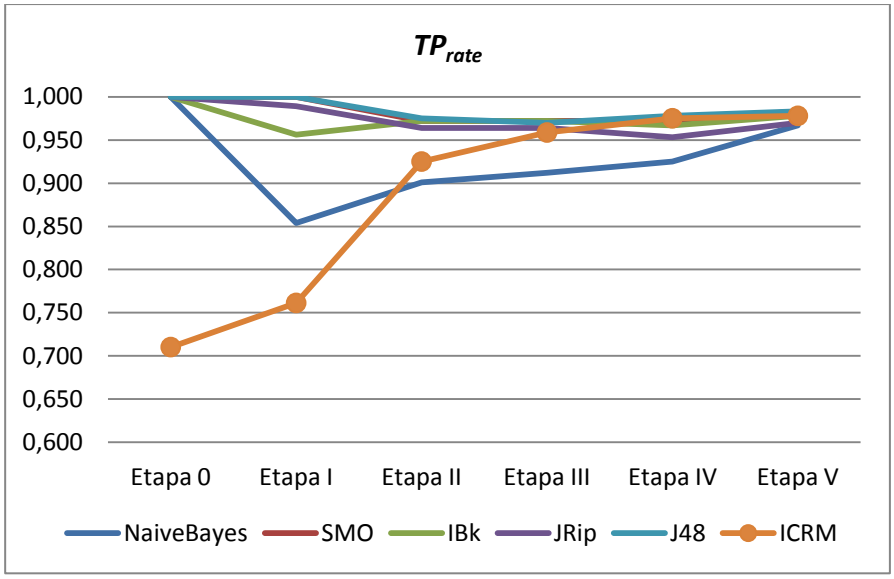


Figura 4.8 TP_{rate} Experimento 2.

En la Etapa I sólo seis atributos sobre aspectos escolares y algunas condiciones sociales del estudiante fueron integrados al conjunto de datos como atributos más relevantes. Los resultados de clasificación obtenidos por las cuatro medidas usadas fueron similares a los obtenidos en el primer experimento y nuevamente el algoritmo ICRM fue el mejor.

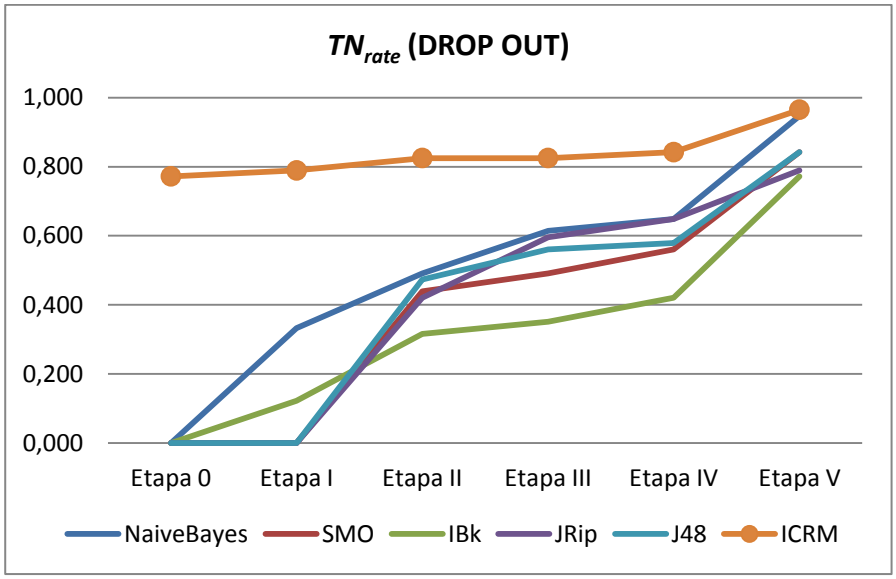


Figura 4.9 TN_{rate} Experimento 2.

En la Etapa II sólo tres atributos relacionados con algunas capacidades físicas y hábitos de los estudiantes fueron considerados como los más relevantes para la clasificación. Como puede verse, hay un incremento en el valor de TN_{rate} en todos los algoritmos cuando se agregan al conjunto de datos estos tres atributos y el incremento de TP_{rate} es especialmente notorio en el algoritmo ICRM. Por ello, en esta etapa se puede recomendar el uso del algoritmo ICRM para realizar una predicción lo suficientemente confiable de los alumnos que potencialmente pueden reprobar o abandonar.

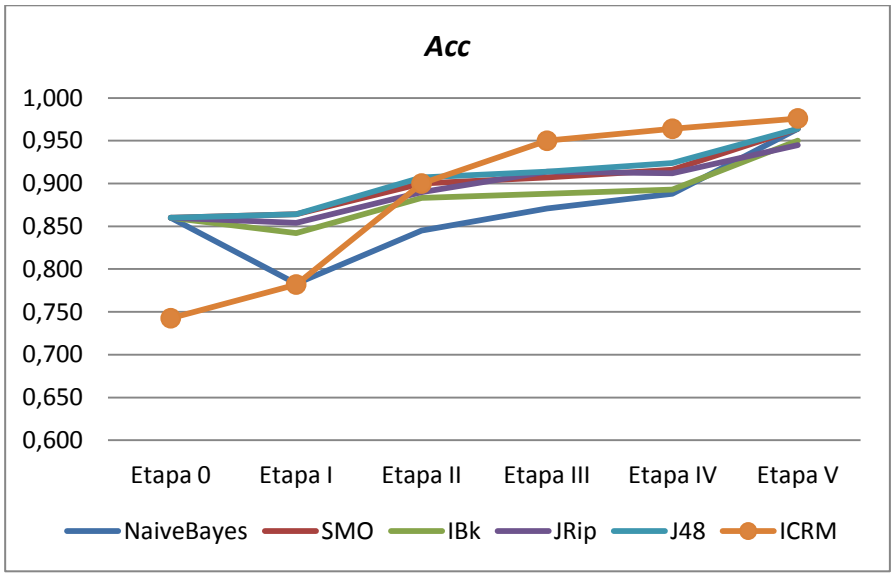


Figura 4.10 *Acc* Experimento 2.

En la Etapa III sólo dos atributos sobre la asistencia y sobre si el estudiante ha tenido algún tipo de sanción administrativa fueron considerados relevantes. Todos los algoritmos mejoraron un poco en las medidas de clasificación, especialmente TP_{rate} y *Acc* del algoritmo ICRM. Así, aunque en esta etapa ya se puede realizar una predicción lo suficientemente confiable de cuáles son los estudiantes que reprobarán o abandonarán, es preferible realizar la predicción desde la etapa previa debido a que la precisión en las medidas de clasificación de los algoritmos es muy similar, si acaso es sólo un poco menor pero lo más importante es que es más temprano.

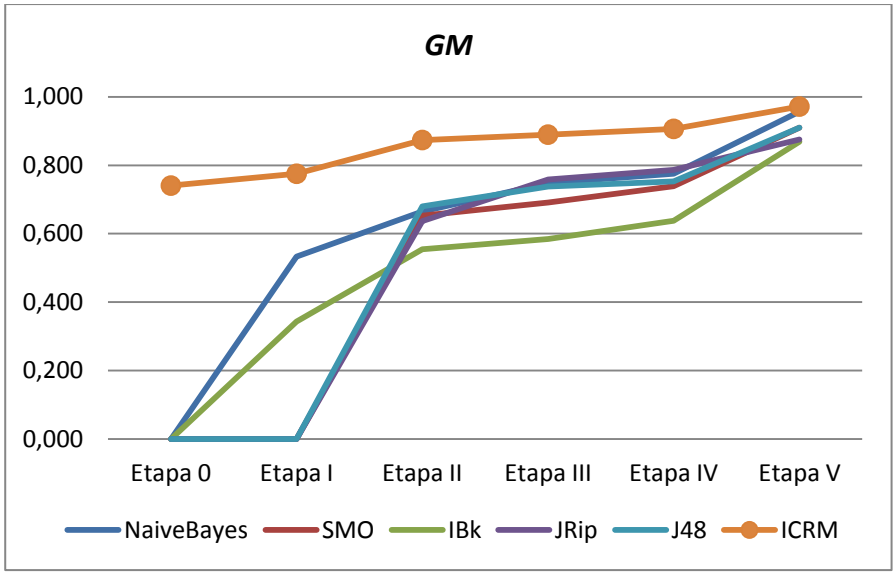


Figura 4.11 *GM* Experimento 2.

En la Etapa IV nuevamente sólo dos atributos, con información sobre si el estudiante cuenta con un espacio adecuado para estudiar y sobre su motivación por estudiar, fueron seleccionados como mejores e integrados al conjunto de datos. Es interesante observar que cuando se agregaron estos atributos, las medidas de clasificación mejoraron aunque muy poco, a diferencia de lo que sucedió en el primer experimento, en donde, en esta etapa sucede lo contrario, es decir, las medidas de clasificación empeoraron un poco. Por otro lado, aunque los valores de las medidas de clasificación en esta etapa ya son suficientemente confiables, el valor máximo de la medida de mayor interés es TN_{rate} = 0.84, conseguido por ICRM, sin embargo, puede considerarse un poco tarde esta etapa como para realizar una predicción temprana ya que se presenta un después de la mitad del periodo escolar.

Finalmente en la última Etapa sólo tres atributos: Notas en Matemáticas 1, Ciencias Sociales 1 y Humanidades 1, fueron seleccionados como mejores. De manera similar al primer experimento, casi todos los algoritmos son capaces de hacer una buena predicción del abandono o reprobación de los estudiantes, con un valor conseguido en la medida TN_{rate} cercano a uno.

4.4 MODELOS DESCUBIERTOS

En este aparto se muestran un par de ejemplos de los diferentes modelos descubiertos por el algoritmo ICRM en cada experimento. El objetivo es analizar su comprensibilidad y utilidad para proporcionar información sobre los estudiantes en riesgo de abandonar o reprobar. Específicamente se muestran los modelos descubiertos en la Etapa II usando los mejores atributos y en la última etapa usando todos los atributos. De esta forma podemos comparar las reglas obtenidas en una predicción temprana frente a las reglas obtenidas por el método tradicional que usa toda la información disponible tras finalizar el curso.

4.4.1 CLASIFICACIÓN EN LA ETAPA II USANDO LOS MEJORES ATRIBUTOS

El siguiente modelo fue obtenido como salida de clasificación del algoritmo ICRM cuando fue aplicado en la Etapa II y usando solamente los mejores atributos:

Reglas de la clase "*REPROBÓ* ":

1. SI (Promedio General en Secundaria < 8 Y Grupo NO ES {1C, 1K, 1O} Y Estudios Madre NO ES Postgrado Y Trabaja > 4 horas) ENTONCES *´REPROBÓ´*.

2. SI (Consumo de Alcohol ES {a menudo, usualmente} Y Fumas ES Si) ENTONCES *´REPROBÓ´*.

3. SI (Número de compañeros ES mayor de 40) ENTONCES *´REPROBÓ´*.

Reglas de la clase "*APROBÓ*":

1. SI (Edad NO ES Mayor de 15 Y Consumo de Alcohol ES {Nunca, Raramente} Y Grupo NO ES {1A2, 1S, 1B2} ENTONCES *´APROBÓ´*.

2. SI (Promedio General en Secundaria > 7.9 Y Estudios madre MAYOR QUE Primaria Y Número de compañeros ES Menos de 30) ENTONCES *´APROBÓ´*.

3. SI (Trabaja < 4 horas Y Fumas ES No) ENTONCES *´APROBÓ´*.

Resultados de las medidas de Clasificación:
Matriz de Confusión.

Actual vs Predicción	APROBÓ	REPROBÓ
APROBÓ	335	27
REPROBÓ	10	47

Acc = 0.91; *GM* = 0.87

Predicciones correctas por clase.
Clase ´*APROBÓ*´: 0.92
Clase ´*REPROBÓ*´: 0.82

Figura 4.12 Salida de ICRM en la Etapa II usando los mejores atributos.

Se puede observar que de las seis reglas obtenidas del tipo *SI-ENTONCES*, tres eran referentes a la clase *REPROBÓ* y tres a la clase *APROBÓ*. Al analizar estas reglas se puede ver que hay una clara relación entre los hábitos de los estudiantes, su condición social y su estatus/rendimiento al final del periodo escolar. Concretamente son indicadores de estudiantes en riesgo de REPROBAR o ABANDONAR: el tener un promedio general en la secundaria menor a ocho, un bajo nivel educativo de la madre, el tiempo dedicado a trabajar de más de 4 horas al día, el consumo regular de bebidas alcohólicas, el fumar y estar en un grupo con más de 40 estudiantes. Por otro lado, cuando la edad no es mayor de 15 años, permite detectar rápidamente a los estudiantes que continuarán sus estudios el siguiente semestre, es decir, aquéllos que APROBARÁN. Los profesores pueden comprobar fácilmente las condiciones mencionadas en la Etapa II del curso, para detectar oportunamente a los estudiantes en riesgo potencial de ABANDONAR o REPROBAR, y así poder ofrecerles algún tipo de ayuda o soporte para intentar evitarlo. Respecto a las medidas de clasificación obtenidas se puede ver que tiene valores altos, siendo mayores a 0.80 y por tanto, el modelo se considera confiable para clasificar a los estudiantes en esta temprana etapa del curso.

4.4.2 CLASIFICACIÓN EN LA ETAPA V USANDO TODOS LOS ATRIBUTOS

El siguiente modelo fue obtenido por el algoritmo ICRM al utilizar la información de la etapa final del curso y usando todos los atributos disponibles.

Reglas de la clase "*REPROBÓ*":

1. SI (Nota en Matemáticas 1 ES ´Pobre´ Y Nota en Computación 1 ES MENOR ´Muy bien´ Y Nota en Inglés 1 ES MENOR ´Excelente´) ENTONCES *´REPROBÓ´*.

2. SI (Nota en Ciencias Sociales 1 ES MENOR ´Regular´ Y Nota en Física 1 ES MENOR ´Bien´) ENTONCES *´REPROBÓ´*.

3. SI (Nota en Taller de Lectura y Redacción 1 ES MENOR ´Regular´) ENTONCES *´REPROBÓ´*.

4. SI (Nota en Humanidades 1 ES MENOR ´Regular´ Y Nota en Física 1 ES ´Pobre´) ENTONCES *´REPROBO´*.

Reglas de la clase "*APROBÓ*":

1. SI (Alcohol ES {Nunca, Raramente} Y Nota en Ciencias Sociales 1 NO ES ´Pobre´ Y Nota en Humanidades 1 NO ES ´Pobre´) ENTONCES *´APROBÓ´*.

2. SI (Asistencia ES ´Buena´ Y Nota en Matemáticas 1 NO ES ´Pobre´ Y Nota en Computación 1 NO ES ´Pobre´) ENTONCES *´APROBÓ´*.

3. SI (Nota en Taller de Lectura y Redacción 1 NO ES ´Pobre´ Y Nivel de Motivación ES ´Aprobaré´) ENTONCES *´APROBÓ´*.

4. SI (Nota en Inglés 1 NO ES ´Pobre´ Y Nota en Física 1 NO ES ´Pobre´) ENTONCES *´APROBÓ´*.

Resultados de las medidas de Clasificación:
Matriz de Confusión.

Actual vs Predicción	APROBÓ	REPROBÓ
APROBÓ	363	0
REPROBÓ	1	56

Acc = 0.99; *GM* = 0.98

Predicciones correctas por clase.
Clase *´APROBÓ´*: 1.00
Clase *´REPROBÓ´*: 0.98

Figura 4.13 Salida de ICRM al final del curso usando todos los atributos.

Se puede ver que de estas ocho reglas del tipo *SI-ENTONCES*, cuatro son sobre la clase *REPROBÓ* y cuatro sobre la clase *APROBÓ*. Al analizar las reglas de la clase *REPROBÓ*, se observa que el obtener malas notas en las asignaturas del curso (Matemáticas 1, Computación 1, Inglés 1, Ciencias Sociales 1, Física 1, Taller de Lectura y Redacción 1 y Humanidades 1) son los únicos factores que aparecen en las reglas y por tanto responsables del fracaso de los estudiantes. Sin embargo, es interesante ver que otros indicadores aparecen en las reglas que detectan a los estudiantes que aprobaron y continuarán el siguiente semestre, como por ejemplo: la abstinencia al consumo de bebidas alcohólicas o que sea muy raro su consumo, tener una buena asistencia a clases y unas altas expectativas de aprobar el semestre. Sobre los valores de las medidas de clasificación del modelo, se puede ver que los valores obtenidos son muy altos, cercanos al máximo posible, es decir al 100%. Sin embargo, este modelo de clasificación no es útil para hacer una predicción temprana, ya que utiliza información que se obtiene al final del semestre, cuando ya no hay tiempo para realizar algún tipo de intervención que permita apoyar a los estudiantes en riesgo de fracasar.

4.5 CONCLUSIONES DEL CAPÍTULO

En este capítulo se ha propuesto una metodología basada en clasificación para predecir lo más temprano posible a los estudiantes en riesgo de reprobar o abandonar la escuela. Se realizaron dos experimentos usando información de 419 estudiantes de la UPUAZ del primer año de bachillerato en México y las principales conclusiones obtenidas son:

- Ha sido posible obtener modelos de clasificación suficientemente confiables para realizar una predicción temprana y oportuna (antes de la mitad del periodo escolar) de los estudiantes que están en riesgo de fracasar. De hecho, se obtuvieron buenos resultados de predicción en las Etapas II y III, es decir, a la cuarta y sexta semana (respectivamente) del inicio del curso. Por tanto, los modelos obtenidos pueden usarse para detección temprana de los estudiantes en riesgo de fracasar. Los profesores y responsables correspondientes pueden tomar conciencia de los estudiantes en riesgo que

tienen y con ello su pueden tomar diferentes medidas para tratar de evitar el fracaso.

- Ha sido posible reducir el número de atributos usado en cada etapa utilizando una técnica de selección de atributos para predecir el abandono y reprobación de los estudiantes. Se obtuvo un valor alto en la medida de la predicción de los estudiantes que reprobaron o abandonaron utilizando conjuntos de datos reducidos en todas las etapas del curso, los cuales tienen solamente los mejores atributos. Concretamente, el modelo de clasificación del algoritmo ICRM en la Etapa II cuando utiliza solamente los diez mejores atributos, obtuvo la precisión suficiente para poder realizar una predicción temprana de los alumnos que fracasaran. Además, estos valores son prácticamente iguales a los obtenidos en la Etapa III que es más tardía y que usa veintisiete atributos. Este hecho es muy importante para el problema que se está tratando, porque permite ahorrar tiempo y también se reduce la cantidad de información que se necesita recoger, es decir, se puede hacer más eficiente todo el proceso de detección.

- El algoritmo propuesto, ICRM, obtuvo en todas las etapas del curso los mejores resultados de clasificación para la predicción de los estudiantes en riesgo de fracasar. Este algoritmo superó a todos los demás algoritmos tradicionales usados, no solamente en la medida TN_{rate}, también en la medida *GM*, la cual indica cuán balanceada es la clasificación de las dos salidas de clasificación, en este caso *APROBÓ* y *REPROBÓ*. Además, el algoritmo ICRM proporciona un modelo de caja blanca, es decir, una salida con reglas muy fácilmente interpretables incluso para un usuario no experto en DM. Las reglas del tipo *SI-ENTONCES* obtenidas, muestran los valores de los atributos que provocan que los estudiantes continúen o no en la escuela y pueden ser utilizados para la toma de decisiones sobre el tipo de apoyo o soporte que se puede brindar a los estudiantes en riesgo, lo cual es uno de los objetivos de los sistemas de alerta temprana.

Finalmente, es importante mencionar que identificar a los estudiantes en riesgo de fracasar por medio de un sistema de alerta temprana es sólo el primer paso para verdaderamente abordar el grave y multifactorial problema del fracaso escolar de los estudiantes. El paso siguiente es identificar las necesidades y problemas específicos

de cada estudiante en riesgo y luego implementar programas apropiados y estrategias para tratar de reducir el fracaso de los estudiantes. Por lo tanto, las partes involucradas deben estar dispuestas a atender las necesidades de los estudiantes en riesgo, para apoyarlos a tiempo y así evitar el fracaso. Por ejemplo, algunas posibles respuestas a las señales de alerta temprana son involucrar a los padres de familia, crear equipos de apoyo académico multi-disciplinares para brindar asesorías con planes de acción individual, realizar seguimientos individuales en los que pueda haber algún tipo de sanción en caso de no atender los apoyos institucionales que se brinden, canalizar al programa de tutorías, etc.

5. CONCLUSIONES Y TRABAJO A FUTURO

En este capítulo se presentan los comentarios finales en forma de conclusiones del trabajo realizado, enumeración de las publicaciones tanto presentadas en congresos como las publicadas en revistas nacionales e internacionales, así como la descripción de las futuras líneas de investigación que se proponen como continuación del trabajo presentado en esta memoria de tesis.

5.1 CONCLUSIONES

Las principales conclusiones obtenidas tras el desarrollo del trabajo realizado en esta tesis son las siguientes:

1. Tras hacer una búsqueda o revisión bibliográfica de la literatura relacionada con la predicción del fracaso escolar de los estudiantes, el abandono escolar y la reprobación en diferentes niveles de educación, se ha encontrado una enorme cantidad de trabajos. Esto nos indica que el problema que se trata de resolver en esta tesis es de actualidad y de gran importancia. De hecho el reducir los índices de reprobación, abandono y fracaso es una intención generalizada y creciente de las instituciones educativas de todo el mundo, ya que repercute no solamente en los propios estudiantes que fracasan o suspenden, sino también en sus familias, en las escuelas y en toda la sociedad en general.

2. La tarea de predecir el fracaso escolar de los estudiantes es una tarea muy difícil de conseguir, principalmente por dos causas: la primera es que son muchos los factores que pueden influir en los estudiantes para que reprueben o abandonen sus estudios; y la segunda es que generalmente la información con la que se trabaja para predecir a estos estudiantes está desbalanceada, es decir, no hay igual número de alumnos que aprueban y pasan curso que de alumnos que suspenden, no pasan y/o abandona el curso. Para solventar estas dos dificultades en esta tesis se ha propuesto la utilización de técnicas de DM que desde hace un tiempo atrás se han empezado a utilizar de manera creciente y con éxito, en lugar de las tradicionales técnicas estadísticas.

3. La primera metodología propuesta para predecir a los estudiantes en riesgo de fracaso está basada en la utilización de diferentes técnicas de DM. Durante el pre-procesado se ha realizado un análisis y selección de los mejores atributos para reducir la alta dimensionalidad de los datos. Y se han obtenido mejores resultados usando solamente los mejores atributos en lugar de todos los disponibles. La predicción se ha realizado mediante la utilización de algoritmos de clasificación ya que la clase a predecir era de tipo nominal o categórica. Este método ha obtenido muy buenos resultados de clasificación

con un conjunto de datos reales que estaban desbalanceados. Además, la clase minoritaria, que corresponde a los estudiantes en riesgo, es la que precisamente más interesa en este trabajo. Para ello se han utilizado técnicas de rebalanceo de datos y de clasificación sensible a costos.

4. La segunda metodología propuesta para predecir lo antes posible a los estudiantes en riego de fracasar está basada en aplicar la anterior metodología pero en cada etapa donde se producen los datos en lugar de esperar hasta el final del curso. Utilizando esta metodología se obtienen resultados confiables desde las etapas intermedias en las que estaba dividió el periodo escolar. Concretamente, la predicción de los estudiantes en riesgo se puede hacer en la cuarta semana del inicio del periodo escolar con un buen porcentaje de acierto o exactitud. De esta forma, la información proporcionada por los modelos de clasificación obtenidos se pueden utilizar en un SIAT para tratar de ayudar a los alumnos detectados en riesgo e intentar evitarlo.

5. En la literatura publicada no hay un consenso respecto a que algoritmo de clasificación es el mejor para predecir a los estudiantes en riesgo de fracaso. En algunos casos unos algoritmos obtienen los mejores resultados, pero en cambio en otros casos son otros. Por este motivo, se ha propuesto en lugar de utilizar los algoritmos clásicamente utilizados en este problema, la utilización de un nuevo algoritmo denominado ICRM. Con el algoritmo propuesto se ha obtenido muy buen rendimiento, concretamente los valores más altos en las medidas de evaluación de la clasificación que más importan en nuestro caso para la predicción del fracaso escolar de los estudiantes que son el TN_{rate} y *GM*. Además, el modelo de salida que produce el algoritmo está basado en reglas del tipo *SI – ENTONCES*, las cuales son muy fácilmente interpretables.

6. Este trabajo explica cómo se puede predecir a los estudiantes en riesgo de reprobar o abandonar sus estudios. Ya se mostró que esto se consigue fundamentalmente a través de recoger información de los alumnos, sin embargo, es imposible desarrollar un método infalible, que pueda predecir a la totalidad de los estudiantes que fracasan; siempre hay situaciones que no se pueden predecir y que pueden llevar a los estudiantes a su fracaso. Por otro lado, también se pueden dar los casos en los que los estudiantes en riesgo son

detectados, ellos lo saben, la escuela les ofrece el apoyo que requieren para evitar su fracaso y no están interesados en permanecer en la escuela y simplemente dejan que termine el periodo escolar para no regresar más. Lo anterior suele presentarse en el nivel medio superior de educación, probablemente por la etapa que viven estos estudiantes, los cuales, en su mayoría tienen entre quince y dieciocho años de edad. Así entonces, es imposible terminar con el fracaso escolar de los estudiantes, sin embargo, trabajos como éste, tributan a tratar de reducir el problema.

5.2 DIVULGACIONES

Con el propósito de divulgar los distintos resultados obtenidos de la investigación realizada en esta tesis, se han generado varias publicaciones, las cuales han sido presentadas como artículos de (4) revistas y (1) ponencia en diferentes foros científicos internacionales. A continuación se lista cada uno de ellos agrupados por categorías.

5.2.1 ARTÍCULOS EN REVISTAS INDEXADAS EN EL *JOURNAL CITATION REPORTS DE THOMSON REUTERS*

1. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. C. Marquez-Vera, A. Cano, C. Romero, S. Ventura. Applied Intelligence. 2013. Vol. 38. N. 3, Pag. 315-330.

2. Early Dropout Prediction using Data Mining: A Case Study with High School Students. C. Márquez-Vera, A. Cano, C. Romero, A. S. A. AL-Ghamdic, H. M. Fardounc, S. Ventura. Expert Systems: The journal of Knowledge Engineering. 2015. En 3ª ronda de revision con cambios menores.

5.2.2 ARTÍCULOS EN OTRAS REVISTAS INTERNACIONALES

1. Predicting School Failure and Dropout by Using Data Mining Techniques. C. Marquez-Vera, C. Romero, S. Ventura. IEEE Journal of Latin American Learning Technology. 2013. Vol. 8. N. 1. Pag. 7-14.

5.2.3 ARTÍCULOS EN OTRAS REVISTAS NACIONALES

1. Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. C.M. Vera, C. Romero, S. Ventura. IEEE Rita: Revista Iberoamericana de Tecnologías del Aprendizaje. ISSN: 1932-8540. Vol. 7. Num 3. Nov 2012. Pag. 109-117.

5.2.4 PONENCIAS EN CONGRESOS INTERNACIONALES INDEXADOS EN CORE

1. Predicting School Failure Using Data Mining. C. Marquez-Vera, C. Romero, S. Ventura. International Conference on Educational Data Mining. Eindhoven, Holland, 6-8 July, 2011. Pag. 271-275.

5.3 TRABAJO FUTURO

Como líneas de trabajo futuro y que sirven de continuación y mejora de esta tesis, se plantean las siguientes tareas:

- Aplicar los dos modelos propuestos de predicción del fracaso escolar y el algoritmo ICRM a datos procedentes de estudiantes de otras partes del mundo y a otros niveles educativos. El objetivo es realizar más pruebas con datos diferentes para poder generalizar los resultados obtenidos, ya que en esta tesis sólo se han aplicado a estudiantes de México que cursaban enseñanzas medias. Para ello habría que comprobar si se obtienen los mismos buenos resultados de predicción con otro tipo distinto de estudiantes. También habría que estudiar y analizar los resultados de clasificación que se obtienen si no se

utilizan exactamente los mismos factores o atributos que se han utilizado en esta tesis, ya sea porque no se disponga de todos, o ya sea porque se disponga de otros completamente nuevos atributos y específicos de un nivel de educación, o ya sea porque se obtienen en un orden distinto de las etapas del curso.

- Desarrollar una herramienta software específica que esté orientada para ser usada por un profesor, coordinador de curso o una persona no experta en DM. El objetivo de la herramienta sería integrar y facilitar todo el proceso de descubrimiento de conocimiento desde el pre-procesado de los datos hasta la visualización de los modelos descubiertos pasando por la ejecución de los algoritmos de clasificación. De esta forma se evitaría tener que hacer un pre-procesado manual, la utilización del software Weka y la ejecución desde línea de comandos del algoritmo ICRM.

- Integrar toda la metodología propuesta dentro de un sistema real de alerta temprana de una institución educativa. Para ello, habría que añadir después de detectar a los estudiantes en riesgo, un conjunto de posibles acciones a realizar como por ejemplo: asesorías académicas, apoyo psicológico, tutorías y seguimiento, dotación de ayudas y becas, involucrar a los familiares, matriculación en cursos de desintoxicación, etc. Además, habría que evaluar cada tipo de intervenciones mencionadas y así poder evaluar el efecto de las diferentes intervenciones y poder determinar cuáles son las más apropiadas para cada tipo de estudiante en riesgo. Por tanto, para poder finalizar y evaluar todo este proceso de forma completa, es necesario disponer de información sobre los resultados obtenidos después de aplicar estas intervenciones a los estudiantes.

REFERENCIAS BIBLIOGRÁFICAS

Aguilar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. En *LAK '14 Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (págs. 103 - 112). New York, USA: ACM.

Aguilar, S., Lon, S., & Teasley, S. D. (2014). Perceptions and Use of an Early Warning System During a Higher Education Transition Program. En *LAK '14 Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (págs. 113 - 117). New York, USA: ACM.

Aha, D., & Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*(6), 37 - 66.

Aloise-Young, P. A., & Chávez, E. L. (2002). Not all school dropouts are the same: Ethnic differences in the relation between reason for leaving school and adolescent substance use. *Psychol Sch., 39*(5), 539 - 547.

Álvarez Aldaco, L. A. (2009). 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*. Veracruz, Veracruz, México.

Antunes, C. (2010). Anticipating students´failure as soon as possible. En *Hanbook of Educational Data Mining.* (págs. 353 - 364). CRC Press.

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing University Drop Out Rates. *Computers & Education, 53*, 563 - 574.

Baker, R., & Yacef, K. (2009). The state of Educational Data Mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

Bayer, J. B., Geryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting dropout from social behaviour of students. *International Conference on Educational Data Mining.*, (págs. 103 - 109).

Braxton, J. M., Jhonson, R. M., & Shaw-Sullivan, A. V. (1997). Appraising Tinto´s Theory of College Student Departure. *Higher Education: Handbook of Theory and Research., 12*. N.Y, USA: Agathon.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Clasification and regression trees.* New York: Chapman & Hall.

Cano, A., Zafra, A., & Ventura, S. (2011). An EP algorithm for learning highly interpretable classifiers. *Proceedings of the 10th international conference on intelligent systems design and applications* (págs. 325 - 330). ISDA´11.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying Data Mining Techniques to e-learning system. En L. Jain, R. Tedman, & D. Tedman, *Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence* (Vol. 62, págs. 183 - 221). Springer-Verlang.

Cendrowska, J. (1987). Prism: an algorithm for inducing modular rules. *Man - Mach Stud, 27*(4), 349 - 370.

Chanyoung, L., & Orazem, P. F. (2010). High School Employment, School Performance and College Entry. *Economics of Education Review, 29*, 29 - 39.

Chawla, N. V., Bowyer, K. W., & Hall, L. O. (2002). Synthetic minority over-sampling technique. *J. ArtifIntell*(16), 321 - 357.

CIET. (1995). *Determinantes de la deserción y repetición escolar en el primero y segundo ciclo.* Recuperado el 27 de 04 de 2014, de Reporte de proyecto CIET: www.ciet.org

Cohen, W. (1995). Fast effective rule induction. *Twelfth international conference on machine learning*, (págs. 115 - 123).

Corral Verdugo, V., & Diáz Núñez, X. (2009). Factores Asociados a la Reprobación de los Estudiantes de la Universidad de Sonora. *X Congreso Nacional de Investigación Educativa.*

Deegalla, S. B. (2006). Reducing high-dimensional data by principal component analysis vs rndom projection for nearest neiighbor classification. *International conference on machine learning and applications*, (págs. 245 - 250).

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Dropo Out: A Case Study. *2nd Int. Conf. On Educational Data Mining.*

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention mangment. *Decision Support Systems, 49*, 498 - 506.

Diosan, L., Rogozan, A., & Pecuchet, J. (2012). Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters. *Appl Intell*(36), 280 - 294.

Elkan, C. (2001). The foundations of cost-sensitive learning. *International joint conference on artificial intelligence.*, (págs. 1 - 6).

Escudero Muñoz, J. M., González González, M. T., & Martínez Domínguez, B. (2009). El Fracaso Escolar como Exclusión Educativa: Comprensión, Políticas y Prácticas. *Revista Iberoamericana de Educación*, Vol: 50. 41 - 64.

Esparrells, C. P. (2004). La Educación Universitaria en España: El vínculo entre financiación y calidad. *Revista en Educación*, 305 - 316.

Espíndola, E., & León, A. (2002). La Deserción Escolar en América Látina un Tema Prioritario para la Agenda Regional. *Revista Iberoamericana de Educación.*(30), 1 - 17.

Fourtin, L., Marcotte, D., Potvin, P., Roger, E., & Joly, J. (2006). Typology of students at risk of dropping out of school: description by personal, family and school factors. *Psychol Educ*, *21*(4), 363 - 383.

Freund, Y., & Mason, L. (1999). The alternating decision tree algorithm. *Proceedings of the 16th international conference on machine learning*, (págs. 124 - 133).

Giménez, P. (2005). *Entorno Social.* Recuperado el 28 de 4 de 2014, de Absentismo Escolar: Causas y Soluciones al Fracaso Escolar: www.entornosocial.es

González González, M. T. (2006). Absentismo y Abandono Escolar: Una situación Singular de la Exclusión Educativa. *Revista electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, Vol. 4. 1 - 15.

Grasso, V. F. (2012). *Early Warning Systems: State of Art Analysis and Future Directions.* United Nations Environment Programme (UNEP)., Division of Early Warning and Assessment (DEWA)., Nairobi.

Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data mining on imbalanced data sets. *Proceedings of intenational conference on advanced computer theory and engineering.*, (págs. 1020 - 1024).

Hall, M. A., & Holmes, G. (2002). *Benchmarking attribute slection techniques for data mining.* Technic, University of Waikato, Computer Science, Waikato.

Hämäläinen, W., & Vinni, M. (2011). Classifiers for educational data mining. En *Handobook of Educational Data Mining.* London.: Chapman & Hall/CRC.

Heiner, C., Baker, R., & Yacef, K. (2006). Proceedings of the workshop on Educational Data Mining. *8th International Conference on Intelligent Tutoring Systems.*

Heppen, J., & Bowles, S. (2008). *Developing Early Warning Systems to identify potential high school dropouts.* Retrieved from betterhighschools.org.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Mach Learn*(11), 63 - 91.

Inan, F. A., Yukselturk, E., & Grant, M. (2006). Profiling Potential Dropout students By Individual Characteristics in an Online Certificate Program. *International Journal of Instructional Media, 36*(2), 1 - 8.

Jiménez, M., Luna, J. M., & Ventura, S. (2013). EDM para la detección precoz del fracaso escolar en secundaria., (págs. 1353 - 1362). Madrid.

John, G., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, (págs. 338 - 345). San Mateo.

Klösgen, W., & Zytkow, J. M. (2002). *Handobook of data mining and knowledge discovery.* London: Oxford University Press.

Koedel, C. (2008). Teacher Quality and Dropout Outcomes in a Large Urban School District. *Journal of Urban Economics, 6*, 560 - 572.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on artificial intelligence*, (págs. 1137 - 1143).

Kotsiantis, S. B. (2009). Educational Data Mining: A case Study for Predicting Dropout Prone Students. *Int. J. Knowledge Engineering and Soft Data Paradigms, 1*(2), 101 - 111.

Kotsiantis, S. B. (2012). Use of machine learning technics for educational proposes: a decision support system for forcasting students´grades. *Artif Intell Rev.*, 331 - 344.

Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting students´marks in Hellenic Open Univesity. *IEEE International conference on advanced learning technologies*, (págs. 664 - 668).

Kotsiantis, S., Patriarchas, K., & Xenos, M. (2010). A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´Performance in Distance Education. *Knowledge-Based System.*, 529 - 535.

Kovacic, Z. J. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. *Informing Science & IT Education Conference.*, (págs. 647 - 665).

Levy, Y. (2007). Comparing Dropouts and Persistence in e-learning Courses. *Computer & Education*(48), 185 - 204.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computer and Education*(53), 950 - 965.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning System" for educators: A proof of concept. *Computer & Education., 54*(2), 588 - 599.

Magaña Hernández, M. (2002). Causas del Fracaso Escolar. *XIII Congreso de la Sociedad Española de Medicina Adolescente*. España.

Maldonado-Ulloa, P. Y., Sancén-Rodríguez, A. J., Torres-Valadés, M., & Murillo-Pazarán, B. (2011). *Programa Síguele. Sistema de Alerta Temprana. Lineamientos de Operación.* (S. d. México., Productor) Obtenido de SEMS.

Márquez-Vera, C., Romero, C., & Ventura, S. (2011). Predicting school failure using data mining. *4th Educational Data Mining Conference.*, (págs. 271 - 276). Eindhoven.

Martínez, D. (2001). Predicting student outcomes using discriminant functions analysis. En *Annual meeting of the research and planning group.* (págs. 163 - 173). California.

Más-Estellés, J., Alcover-Arándiga, R., Dapena-Janeiro, A., Valderruten-Vidal, A., Satorre-Cuerda, F., Llopis-Pascual, F., y otros. (2009). Rendimiento Académico de los Estudios de Informática en Algunos Centros Españoles. *XV Jornadas de Enseñanza Universitaria de la Informática.* Barcelona.

Massa, S., & Puliafito, P. (1999). An Application of Data Mining to the Problem of Data Mining to the Problem of the University Students´ Dropout Using Markov Chains. In *Principles of Data Mining and Knowledge Discovery. Lectures Notes in Computer Science* (Vol. 1704, pp. 51 - 60). J. M. Zyttkow and Rauch.

McDonald, B. (2004). Predicting Student Success. *Journal for Mathematics Teaching and Learning.*(1), 1 - 14.

Mellalieu, P. J. (2011). Predicting success, excellence, and retention from students´early course performance: progress results from a data mining based

decision support system in a first year tertiary education programme. *International Conference of the International Council for Higher Education.*, (págs. 1 - 9). Florida. USA.

Méndez, G., Buskirk, T. D., Lohor, S., & Haag, S. (2008). Factors associated with persistence in science and enginering majors: an exploratory study using classification trees and random forests. *Journal of Engineering Education., 97*, 57 - 70.

Monge Reyes, M., & Martínez Godínez, B. (5 de 6 de 2006). *Articulos Gratis.* Recuperado el 28 de 04 de 2014, de La Reprobación Escolar un Fenómeno Latente en el Sistema Educativo Actual: www.articulosgratis.com

Moseley, L., & Mead, D. (2008). Predicting who will drop out of nursing courses: a machine learning exercise. *Nurse Educ Today*(28), 363 - 383.

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational Leadership., 65*(2), 28 - 33.

Pan, W. (2012). The use of genetic programming for the construction of a fianancial management model in an enterprise. *Appl Intell*(36), 271 - 279.

Parker, A. (1999). A study of variables that predict dropout from distance education. *Int. J. Educ. and Technol., 1*(2), 1 - 11.

Parker, A. (2003). Identifying Predictors of Academic Persistence In Distance Education. *USDLA Journal*, 1 - 8.

Platt, J. (1998). *Fast Training of Support Vector Machines using Sequential Minimal Optimisation.* B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning.

Quadril, M. N., & Kalyankar, N. V. (2010). Drop Out Feature os student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology, 10*, 2 - 5.

Quinlan, J. (1983). *C45. Programs for machine learning.* San Mateo: Morgan Kaufman.

Reyes-Seánez, M. (2006). Una Reflexión Sobre la Reprobación Escolar en la Educación Escolar como Fenómeno Social. *Revista Iberoamericana de Educación., 39*(7), 1 - 6.

Richards, D. (2009). Two decades of RDR research. *Knowl Eng Rev, 24*(2), 159 - 184.

Rodallegas Ramos, E., Torres González, A., Gaona Couto, B. B., Gastelloú Hernández, E., Lezama Morales, R. A., & Valero Orea, S. (2010). Modelo Predictivo para la Determinación de Causas de Reprobación Escolar Mediante Minéría de Datos. En M. E. Prieto, J. M. Dodero, & D. O. Villegas, *Recursos Digitales para la Educación y la Cultura, Volumen KAAMBAL* (págs. 48 - 55). Yucatan, México: UTM - UCA.

Rogers, T., Colvin, C., & Chiera, B. (2014). Modest analytics: using the index method to identify students at risk of failure. *LAK '14 Proceedings of the Fourth International Conference on Learning Analytics And Knowledge.* (págs. 118 - 122). New York: ACM.

Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert System with Applications*(33), 135 - 146.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*(6), 601 - 618.

Romero, C., & Ventura, S. (2013). Data mining education. *WIREs Data Mining Knowledge Discovery., 3*, 12 - 27.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of MOODLE students. *Computer Applications in Engineering Education Journal., 21*, 135 - 146.

Roy, S. (2002). Nearest neighbor with generalization. *Master´s thesis*. New Zeland: University of Canterbury, Christchurch.

Seidman, A. (1996). Retention revisited: RET = EId+(E+I+C)Iv. *College and University, 71*(4), 18 - 20.

Silas-Casillas, J. C. (2009). Estudiar en la Montaña sin Morir en el Intento. *Revista Internacional de Investigación Educativa, 2*(3), 211 - 226.

Slavin, R. E., Karweit, N. L., & Wasik, B. A. (1994). *Preventing early school failure.* Allyn & Bacon.

SUBSECRETARÍA DE EDUCACIÓN MEDIA SUPERIOR - SEMS. (5 de Noviembre de 2013). *SECRETARÍA DE EDUCACIÓN PÚBLICA*. Recuperado el 29 de Julio de 2014, de SISTEMA DE ALERTA TEMPRANA - SIAT: http://www.sems.gob.mx/en_mx/sems/sistema_alerta_temprana_siat

Subsecretaria de Educación Media Superior SEMS. (26 de 04 de 2014). Recuperado el 26 de 04 de 2014, de Lineamientos de Operación del Fondo para Fortalecer la Autonomía de Gestión en Planteles de Educación Media Superior 2014: www.sems.gob.mx

Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achivement of first-year university students using data mining methods. *Educational data mining workshop.*, (págs. 1 - 8).

Tinto, V. (1975). Dropout from Higher Education: A Theorical Synthesis of Recent Research. *Review of Educational Research*, 89 - 125.

Tinto, V. (1987). *Leaving college: rethinking the causes and cures of students attrition.* Chicago: University of chicago Press.

Uekawa, K., Merola, S., Fernandez, F., & Porowski, A. (2010). Creating an Early Warning System: Predictors of Dropout in Delaware. *Regional Educational Laboratory Mid Atlantic.*, 1 - 50.

Valero Orea, S. (2009). *www.utim.edu.mx.* Recuperado el 25 de 4 de 2014, de http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf

Vassiliou, A. (2013). Early warning systems in Europe: practice, methods and lessons. *Thematic Working Group on Early School Leaving*, 1 - 17.

Veitch, W. (2004). Identifying characteristics of high school dropouts: data mining with a decisión tree model. En *Annual meeting of the American educational Research Association.* (págs. 1 - 11).

Wang, A. Y., & Newlin, M. H. (2002). Predictors of web-based performance: the role of self-efficacy and reasons for taking an on-line class. *Computers in Human Behavior., 18*, 151 - 163.

Wegner, L., Flisher, A. J., Lombard, C., & King, G. (2008). Leisure boredom and high school dropout in Cape Town, South Africa. *Journal of Adolescence., 31*, 421 - 431.

Whigham, P. (1996). Grammatical bias for evolutionary learning. *PhD Dissertation*. University of New South Wales.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining, Practical machine learning tools and techniques.* USA: Morgan Kaufman.

Xenos, M., Pierrakeas, C., & Pintelas, P. (2002). A Survey on Student Dropout Rates and Dropout Causes Concerning the Students in the Course of Informatics of the Hellenic Open University. *Computers & Education*, 361 - 377.

Xenos, Pierrakeas, & Pintelas. (2002). A Survey on Student Dropout Rates and Dropout Causes Concerning the Students in the Course of Informatics of the Hellenic Open University. *Computer & Education*(39), 361 - 377.

Zhang, G. P. (2000). Neural Networks for Classification: A survey. En *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART C: APPLICATIONS AND REVIEWS* (Vol. 30, págs. 451 - 462).

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use data mining to improve student retention in higher education - A case study. *Proceedings of the 12th International Conference on Enterprise Information Systems. 1*, págs. 190 - 197. Madeira, Portugal: ICEIS.

APÉNDICE A: ENCUESTA PARA DETECTAR FACTORES QUE AFECTAN EL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES

En este apartado se incluye la encuesta aplicada a los estudiantes de primer semestre del Programa II de la UAPUAZ, la cual tiene la intención de detectar algunos de los factores que afectan el rendimiento académico de los estudiantes. Este instrumento permitió construir el conjunto de datos de trabajo.

UNIVERSIDAD AUTÓNOMA DE ZACATECAS.
UNIDAD ACADÉMICA PREPARATORIA.

ENCUESTA PARA DETECTAR FACTORES QUE AFECTAN EL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES.

Instrucciones: Responde o completa seriamente lo que se indica rellenando los □.

A. PATERNO	**A. MATERNO**	**NOMBRE**

EDAD (AÑOS)	**SEMESTRE Y GRUPO**	**TURNO**	**SEC. PROCEDENCIA**

1. Al final del semestre Agosto – Diciembre 2010, tú crees que:
   - □ Aprobaré todas mis materias con buenas calificaciones.
   - □ Aprobaré todas mis materias con regulares calificaciones.
   - □ Probablemente repruebe una materia.
   - □ No creo aprobar este semestre.
2. ¿Cuántos amigos tienes en el salón de clases?
   - □ Ninguno.
   - □ 1.
   - □ De 2 a 5.
   - □ Más de 6.
3. ¿Cuánto tiempo dedicas a tus estudios adicionalmente a la jornada diaria escolar?
   - □ Menos de 1 h.
   - □ De 1 a 2 h.
   - □ Más de 2 h.
4. Generalmente estudias…
   - □ Solo.
   - □ En grupo.
5. Generalmente estudias…
   - □ En casa.
   - □ En la biblioteca.
   - □ En casa de un amigo.
6. Estudias…
   - □ Cuando tengo tarea o presentaré examen.
   - □ Aunque no tenga tarea o examen.

7. Cuando tienes dudas sobre algún tema entonces…
   - ☐ Voy con el profesor para que me ayude a aclararlas.
   - ☐ Se las comento a un compañero para que me ayude a aclararlas.
   - ☐ Trato de resolverlas yo mismo.
8. Religión
   - ☐ Católica.
   - ☐ Otra.
   - ☐ Ninguna.
9. La decisión de que carrera universitaria estudiar en su momento…
   - ☐ La tomaré yo.
   - ☐ Mis padres y/o familiares influirán en mi decisión.
10. De tu personalidad o forma de comportarte se puede decir que eres…
    - ☐ Abierto y hablador.
    - ☐ Serio.
    - ☐ Tímido.
11. ¿Tienes alguna discapacidad física o sicológica?
    - ☐ Sí.
    - ☐ No.
12. ¿Tienes o has tenido alguna enfermedad grave?
    - ☐ Sí. ¿Cuál? _____
    - ☐ No.
13. En cuanto al consumo de bebidas alcohólicas…
    - ☐ Las consumo frecuentemente.
    - ☐ Las consumo sólo cuando voy a alguna fiesta o reunión.
    - ☐ Solamente las he probado, pero no me gustan y no las consumo.
    - ☐ Nunca las he probado.
14. ¿Fumas?
    - ☐ Sí.
    - ☐ Ocasionalmente.
    - ☐ No.
15. En cuanto al nivel económico de ingreso en tu familia es…
    - ☐ Alto.
    - ☐ Medio.
    - ☐ Bajo.
    - ☐ Muy bajo.
16. ¿Dispones de dinero para costear todos los gastos que se necesitan para que estés estudiando?
    - ☐ Sí.
    - ☐ No.
17. ¿Dispones de algún tipo de ayuda o beca que te apoye para realizar tus estudios?

- ☐ Sí.
- ☐ No.

18. ¿Trabajas para ayudar a la manutención de tu familia?
    - ☐ Sí. ¿Cuánto tiempo al día? _____
    - ☐ No.
19. Tú vives…
    - ☐ Con mis padres (mamá y papá).
    - ☐ Sólo con uno de mis padres.
    - ☐ Con un familiar.
    - ☐ Solo.
20. Del nivel de estudios de tus padres. Marca con una X:

MADRE		**PADRE**	
	Primaria		Primaria
	Secundaria		Secundaria
	Preparatoria		Preparatoria
	Licenciatura		Licenciatura
	Postgrado		Postgrado
	No sé		No sé

21. ¿Cuántos hermanos tienes?
    - ☐ Ninguno.
    - ☐ 1.
    - ☐ 2.
    - ☐ 3.
    - ☐ 4 o más.
22. Orden de nacimiento que ocupas entre tus hermanos
    - ☐ Primero (soy el mayor).
    - ☐ Segundo.
    - ☐ Tercero.
    - ☐ Otro. Especifica _____
23. ¿En tu vivienda tienes un espacio adecuado (con mesa o escritorio) para estudiar?
    - ☐ Sí.
    - ☐ No.
24. ¿Tus padres están interesados en tus estudios, fomentan que sigas estudiando y te estimulan?
    - ☐ Sí.
    - ☐ No.
25. ¿Cuántos años llevas viviendo en la misma localidad? _____
26. Modo de transporte para ir a la escuela
    - ☐ Automóvil de la familia.
    - ☐ Transporte público.
    - ☐ Andando.

- ☐ Otro. Especifica: _____

27. Distancia de tu domicilio a la escuela en kilómetros
    - ☐ Menos de 2 km.
    - ☐ Entre 2 y 10 km.
    - ☐ Más de 10 km.
28. ¿Asistes todos los días a clases?
    - ☐ Sí.
    - ☐ No.
29. ¿Te aburres en clases?
    - ☐ Sí.
    - ☐ Ocasionalmente.
    - ☐ No.
30. ¿Crees que los conocimientos que estas adquiriendo en la escuela te son de utilidad?
    - ☐ Sí.
    - ☐ Sólo algunos.
    - ☐ No.
31. ¿Alguna asignatura es para ti especialmente difícil?
    - ☐ No.
    - ☐ Sí. ¿Cuál? _____
32. ¿Tomas apuntes o notas mientras el profesor expone su clase?
    - ☐ Sí.
    - ☐ No.
33. ¿Consideras que los profesores te dejan mucho trabajo o tareas para hacer en casa?
    - ☐ Sí.
    - ☐ No.
34. Número de alumnos en tu grupo.
    - ☐ Menos de 20.
    - ☐ Entre 20 y 30.
    - ☐ Entre 30 y 40.
    - ☐ Más de 40.
35. En general la forma de enseñar de tus profesores es…
    - ☐ Buena.
    - ☐ Regular.
    - ☐ Mala.
36. En cuanto a las instalaciones que tiene la escuela (Biblioteca, Laboratorios, Canchas deportivas, Áreas verdes, etc.).
    - ☐ Sí hay, son suficientes y están en buenas condiciones.
    - ☐ Hay pero no son suficientes.
    - ☐ No hay.

37. ¿Tienes algún tipo de Tutor o Asesor escolar que esté pendiente de tus estudios?
    - ☐ Sí.
    - ☐ No.
38. ¿Crees que la institución y sus profesores se preocupan por tu desarrollo escolar?
    - ☐ Sí.
    - ☐ No.
39. Tu calificación promedio en la secundaria fue:
    - ☐ entre 9.0 y 10.
    - ☐ entre 8.0 y 8.9.
    - ☐ entre 7.0 y 7.9.
    - ☐ entre 6.0 y 6.9.
40. En el examen de admisión (EXANI I) a la UAPUAZ tu resultado fue:
    - ☐ Excelente.
    - ☐ Muy bueno.
    - ☐ Bueno.
    - ☐ Regular.
    - ☐ Malo.
    - ☐ No presenté.
41. ¿Has reprobado en alguna ocasión, de manera que te hayas retrasado en algún periodo escolar?
    - ☐ Sí.
    - ☐ No.

42. ¿Has abandonado alguna ocasión tus estudios o has sido expulsado de alguna escuela?
    - ☐ Sí.
    - ☐ No.

Fecha: _____

TÍTULO DE LA TESIS: Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos

DOCTORANDO/A: Carlos Márquez Vera

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

El doctorando (Carlos Máquez Vera) ha progresado enormemente como investigador desde que en el año 2010 realizara su trabajo de investigación tutelada con los mismos directores y temática, que dio pie a la realización de la actual tesis.

Durante estos 5 años el doctorando ha trabajado duro tanto a distancia como realizando varias estancias, y ha seguido siempre las pautas de trabajo que le hemos marcado los directores.

Como fruto del buen trabajo realizado, de esta tesis se han derivado las siguientes publicaciones:

- 1 Artículo publicado en revista indexada en el JCR.
- 1 Artículo actualmente en 3º ronda de revisión en revista indexada en el JCR.
- 1 Artículo en revista internacional con revisión por pares.
- 1 Artículo en revista nacional.
- 1 Artículo en congreso internacional core B.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, _18__ de __Mayo_____ de ___2015_____

Firma del/de los director/es

Fdo.:__Sebastián Ventura Soto____ Fdo.: ___Cristóbal Romero Morales_____