Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los *k* vecinos más cercanos

Sergio Valero Orea[1], Alejandro Salvador Vargas[1], Marcela García Alonso[1]

[1] Universidad Tecnológica de Izúcar de Matamoros, Prolongación Reforma 168, Santiago Mihuacán, 74420, Izúcar de Matamoros, Puebla, México
svalero@utim.edu.mx, salvar73@hotmail.com, mgarcia@utim.edu.mx,

Resumen. Las técnicas de minería de datos permiten obtener conocimiento oculto en grandes cantidades de datos con información valiosa que, al explotarse, ofrece ventajas competitivas a las organizaciones. En el caso de las instituciones de educación superior, existen muchos datos respecto a los estudiantes, útiles para tomar decisiones estratégicas en pro de los mismos. Con base en esto, se han aplicado técnicas de minería de datos para buscar predecir la deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros, tomando como base de análisis los datos del estudio socioeconómico del EXANI-II, elaborado por el CENEVAL, mismo que se aplica desde el 2003 en nuestra institución. Para esta investigación se utilizaron específicamente dos algoritmos: el algoritmo de árboles de clasificación C4.5 y el algoritmo de los *k* vecinos más cercanos.

Palabras clave: Minería de datos, deserción escolar, árboles de decisión, *k* vecinos más cercanos.

1 Introducción

La minería de datos es una subdisciplina de las ciencias de la computación que ha logrado mucho reconocimiento en los últimos años, principalmente porque puede ser usada para diferentes propósitos como herramienta de apoyo en las demás disciplinas de las ciencias. Su fortaleza radica en el hecho de que forma parte del proceso de descubrimiento del conocimiento, cuyo objetivo es la búsqueda de patrones de datos que sean válidos, novedosos, potencialmente útiles y comprensibles [1].

La minería de datos en la educación no es un tópico nuevo y su estudio y aplicación ha sido muy relevante en los últimos años. El uso de estas técnicas permite, entre otras cosas, predecir cualquier fenómeno dentro del ámbito educativo. De esta forma, utilizando las técnicas que nos ofrece la minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de desertar de cualquier estudiante.

2 Antecedentes

De acuerdo con la ANUIES [2], en México de cada 100 estudiantes que ingresan a la Instituciones de Educación Superior (IES), sólo 60 egresan y de éstos, sólo 20 se titulan. De acuerdo con la Organización para la Cooperación y el Desarrollo Económico [3], aproximadamente un tercio de los estudiantes de educación superior en México desertarán antes de completar sus estudios de nivel superior.

Al igual que en muchas IES, la deserción escolar es un grave problema de la Universidad Tecnológica de Izúcar de Matamoros (UTIM). Muchos factores influyen en la deserción, sin embargo, al no haber un diagnóstico oportuno, conlleva a la falta de seguimiento al problema. Considerando un índice de deserción relativamente alto (tabla 1), encontramos un área de oportunidad de poder predecir la posibilidad de deserción de los estudiantes.

Tabla 1. Relación ingreso/deserción por periodo cuatrimestral en la UTIM.

PERIODO	INGRESO	DESERCIÓN	PORCENTAJE
Sep. - Dic. 2004	881	102	11.58%
Ene. - Abr. 2005	779	73	9.37%
May. - Ago. 2005	706	37	5.24%
Sep. - Dic. 2005	742	77	10.38%
Ene. - Abr. 2006	665	55	8.27%
May. - Ago. 2006	610	20	3.28%
Sep. - Dic. 2006	789	78	9.89%
Ene. - Abr. 2007	711	30	4.22%
May. - Ago. 2007	681	30	4.41%
Sep. - Dic. 2007	871	70	8.04%
Ene. - Abr. 2008	801	47	5.87%
May. - Ago. 2008	754	33	4.38%
Sep. - Dic. 2008	1104	68	6.16%
Ene. - Abr. 2009	1036	86	8.30%

En el modelo de Universidades Tecnológicas, existen dos categorías para docentes: Profesores de Tiempo Completo (PTC) y Profesores por Asignatura (PA). Dentro de nuestra institución, los tutores son PTC que guían u orientan a los alumnos durante su estancia en la institución. Estos docentes identifican a los alumnos que desertarán en el momento en que ellos solicitan su baja ya que no existe ningún mecanismo formal que ayude a identificar la vulnerabilidad de los estudiantes, el resultado del proyecto es una herramienta que permite calcular la probabilidad de deserción de cada uno de los estudiantes, por lo que nuestra aportación a la UTIM tiene la finalidad de apoyar el proceso de tutorías con una herramienta útil y práctica.

3 Minería de datos

La minería de datos es entendida como el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, datawarehouses, o cualquier otro medio de almacenamiento de información [4]. La aplicación de algoritmos de minería de datos requiere de actividades previas destinadas a preparar los datos de manera homogénea. Esta primera etapa es también conocida como ETL (Extract, Transform and Load) [5]. Un proceso completo de aplicación de técnicas de minería, mejor conocido como proceso *de descubrimiento del conocimiento en bases de datos* [6] establece a la minería de datos como una etapa del mismo. Dentro de ésta se pueden utilizar diversos algoritmos predictivos como:

Árboles de decisión C4.5: categorizado como aprendizaje basado en similaridades [8], los árboles de decisión son uno de los algoritmos más sencillos y fáciles de implementar y a su vez de los más poderosos. Este algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información (gain ratio) [4], es decir, elige al atributo que mejor clasifica a los datos.

Técnica de los *k* vecinos más cercanos: conocido como algoritmo de aprendizaje basado en instancias, su funcionamiento es muy simple: se almacenan los ejemplos de entrenamiento de datos históricos y cuando se requiere clasificar a un nuevo objeto, se extraen los objetos más parecidos y se usa su clasificación para clasificar al nuevo objeto [7]. Los vecinos más cercanos a una instancia se obtienen, para el caso de los atributos continuos, utilizando la distancia Euclidiana sobre los n posibles atributos. El resultado de la clasificación por medio de este algoritmo puede ser discreto o continuo. En el caso discreto, el resultado de la clasificación es la clase más común de los k vecinos [7] [8].

4 Desarrollo de la investigación

Para el desarrollo del proyecto, seguimos la propuesta hecha por Hernández [9] en la que se marca el proyecto de minería en una serie de fases definidas: integración y recopilación, selección, limpieza y transformación, minería de datos, pruebas y verificación de resultados.

4.1 Fase de integración

Las fuentes de datos con las que se trabajó, fueron las bases de datos del EXANI – II y de los alumnos inscritos, proporcionados por el departamento de servicios escolares de la UTIM, de los alumnos que causaron baja, así como sus causas.

En resumen, fueron 11 archivos DBF de todas las fechas que se aplicaron EXANI desde el 2003 hasta el 2008, 6 archivos XLS de los alumnos inscritos y la

digitalización de todos los memorándums en donde se notificaba la baja del alumno junto con sus causas.

4.2 Fase de selección, limpieza y transformación

La primera acción realizada fue el análisis de los datos que se insertaron en el almacén de datos, se tuvo que llevar a cabo el proceso ETL para seleccionar los datos útiles para la investigación, después llevar a cabo la limpieza y transformación de los mismos para obtener una vista minable que permita construir un modelo de calidad, realizando operaciones de discretización, sumarización, etc.

Para la base de datos del EXANI, se realizó el proceso ETL para 6 generaciones distintas, trabajando con 11 bases de datos que correspondieron a 4 cuestionarios diferentes. La tabla 2 ejemplifica el grado de complejidad del proceso al trabajar con una gran cantidad de datos, de dominios heterogéneos.

Tabla 2. Diferentes atributos usados por el CENEVAL para representar la situación socioeconómica de un estudiante

Año	Trabaja	Hrs. que trabaja	Tipo de trabajo	Tipo de organización	Trabajo que desarrolla	Ingreso personal
2003	Trabaja	Hrs_trab	Tipo_tra		Trab_des	Ing_per
2004	Trab_act	Hrs_trab	Tip_trab	Org_trab	Ocu_trab	Ingr_per
2005	Trab_act	Hrs_trab	Tip_trab	Org_trab	Ocu_trab	Ingr_per
2006	Trab_act	Hrs_trab	Tip_trab	Org_trab	Ocu_trab	Ingr_per
2007	Trab_act	Hrs_trab				
2008	Trab_act	Hrs_trab				Apor

Para el caso de la base de datos de los alumnos inscritos, se limpiaron y transformaron los datos de 6 archivos tipo XLS, que de la misma forma que el caso anterior, tenían formas distintas de almacenar los datos de los alumnos. Por citar algunos ejemplos, el apellido de algún alumno se almacenaba como “PÉREZ”, “Pérez” o “Perez”, o la fecha de su nacimiento, como “17/01/1985” o “17 de Enero de 1985”.

Por último, una vez capturados los datos de los alumnos que causaron baja (matrícula, nombre y generación), se obtuvo una primer vista minable (mediante SQL), con 16 atributos que representan las características de nuestros alumnos; está vista es útil para aplicar las técnicas de minería de datos.

Tabla 3. Concentrado de atributos finales utilizados para el proceso de minería de datos

Atributo	Valores posibles
Sexo	H (hombre), M (mujer)
Edad	<=18, >18
Tipo_Bach	Bachillerato Abierto, BGO, Bachillerato Pedagógico, Bachillerato Tecnológico (CBTis, CBTa), Colegio de Bachilleres, Preparatoria, Profesional Técnico (CONALEP), Otro.

Prom_Bach	Bajo (Menor a 7), Medio (Entre 7 y 9) y Alto (Superior a 9).
Mat_Rep	Ninguna, 3 o menos, 7 o menos, mas de 7.
IntentosPrev	Si, No
ApoyoEco	Si, No
Inglés	Básico, Intermedio, Avanzado
HabEst	Nulo, Bajo, Medio, Alto
Exani	Bajo (<1000 puntos índice CENEVAL), Medio (entre 1000 y 1150) y Alto (Arriba de 1150).
Esc_Padre	No lo sabe, Sin estudios, Primaria, Secundaria, Media superior y Superior.
Esc_Madre	No lo sabe, Sin estudios, Primaria, Secundaria, Media superior y Superior.
IngresoFam	No lo sabe, <$3000, de $3000 a $6000, de $6001 a $9000, de $9001 a $15000, >$15000
Tam_Fam	Núcleo (padres y a lo más 2 hijos), Extensa (padres y más de 2 hijos).
Trabaja	Si, No
Hrs_Trabajo	Menos 10 hrs semanales, Medio tiempo, Tiempo Completo, No trabaja
Baja	Si, No

4.3 Fase de minería de datos

La tarea de minería de datos seleccionada fue la clasificación, utilizando un árbol de decisión mediante el algoritmo C4.5 y el método de aprendizaje basado en vecindad conocido como los *k* vecinos más cercanos (*k nearest neighbors*). Se crearon muchos árboles de prueba y se ejecutó y probó el algoritmo haciendo las operaciones manualmente. También se construyó un segundo modelo con el algoritmo de los *k* vecinos más cercanos y se compararon los resultados verificando el nivel de confiabilidad de ambos modelos, mismos que se presentan a continuación.

4.4 Fase de pruebas y verificación de resultados

En esta fase se generaron los modelos con la ayuda del minero de datos (Weka) [10]. Se realizaron un conjunto de pruebas que se verificaron al momento de crear los modelos.

Por un lado, para la construcción del árbol de decisión, de las 723 instancias (registros) que formaba nuestra vista minable, Weka tomó 477 instancias (66.6%) para construir el modelo y 246 instancias (33.4) para probarlo, con una precisión del 67.07%. Por otro lado, se probó el segundo modelo con la ejecución del algoritmo de los *k* vecinos más cercanos, utilizando el método de entrenamiento de validación cruzada con 10 evaluaciones, y se pudo notar que al establecer el valor de *k* en 50, se obtuvo una precisión del 67.77%, superior a algoritmo C4.5. Este fue el porcentaje mayor de confiabilidad, ya que se probó el modelo con los valores de k en 1, 10, 50 y 100, obteniendo 62.51%, 67.08%, 67.77% y 67.63%, respectivamente.

Se eligió la construcción del modelo usando el algoritmo para árboles de clasificación ya que presenta un nivel de confiabilidad más alto al trabajar con cantidades mayores de datos. Se pudo notar al construir modelos para poco más de 6500 instancias, y el modelo para el árbol de clasificación tuvo una precisión del

98.98%, mientras que el algoritmo de los *k* vecinos más cercanos apenas y superó el 70% como lo podemos observar en la figura 1:

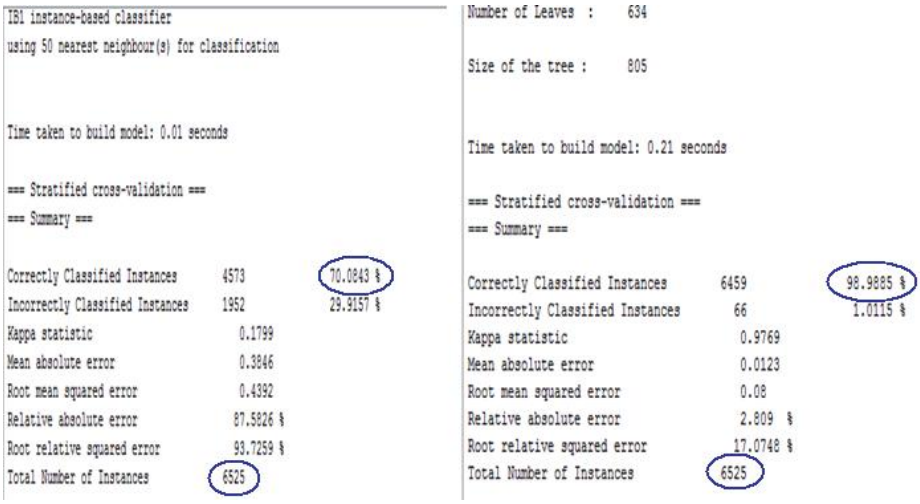

Figura 1. Comparación de resultados: a la izquierda el % de confiabilidad en k vecinos más cercanos y a la derecha el % de confiabilidad de los árboles de decisión

Una vez que se eligió el modelo predictivo con base en árboles de decisión, se procedió a la construcción de una interfaz que permitiera interactuar con el modelo construido. Como se cita en [10], es posible leer una serie de reglas directamente del árbol creado, iniciando en la raíz y recorriéndolo a partir de las decisiones tomadas en cada nodo encontrado hasta llegar a un nodo hoja (nodo final).

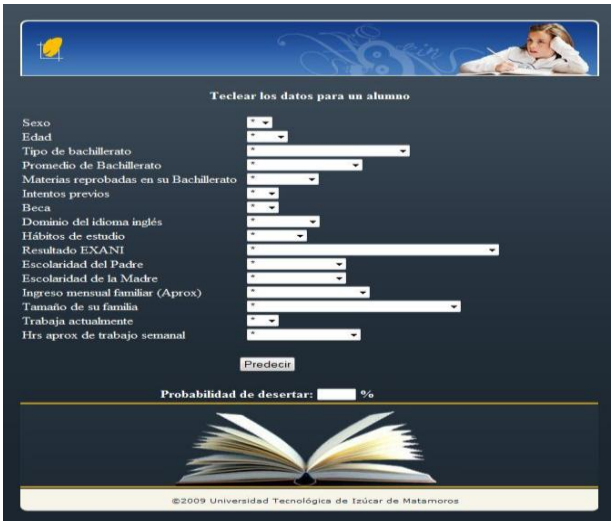

Figura 2. El sistema de predicción de deserción, resultado de la investigación

Usando el sistema de reglas del modelo creado se desarrolló una interfaz Web [11] que facilita la identificación de los alumnos vulnerables. Esta interfaz se encuentra implementada en un Servidor Linux con uso de Apache y PHP.

A través de esta interfaz (figura 2) se probó el modelo con datos de alumnos que se encuentran inscritos en el periodo Septiembre-Diciembre de 2009 del grupo 1A. Este grupo cuenta con 27 alumnos, de los cuales 5 se omitieron por no presentar examen de ingreso, quedando 22 registros para la prueba, cuyos resultados se muestran en la figura 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Nombre	sexo	edad	tipobach	prom	matrep	intentos	beca	ingles	habest	exani	escpadre	escmadre	ingresos	tamfam	trabajas	hrstrab	
2	Alvarez Méndez José Luis	H	19	bgo	6.3	2	no	no	b	b		primaria	primaria	<3000	Nucleo	no	<10	
3	Anacleto Rodríguez Gustavo	H	18	cbta	8.4	3	si	si	i	b	1000	secundaria	secundaria	<3000	Extensa	no	no	90
4	Avila Pelaez Adrián	H	17	cobach	7	1	no	no	b	b	886	media superior	primaria	<3000	Extensa	no	no	100
5	Cano Rosas Juan	H	20	bgo	7.9	0	no	no	b	m	922	secundaria	primaria	<3000	Extensa	no	no	24.6
6	Cortés Gaspar José Juan	H	18	cobach	6.7	4	no	si	i	b		media superior	primaria	<3000	Extensa	no	no	
7	Flores Ruiz Andrés Isai	H	18	cbtis	8.3	0	no	si	i	b	988	primaria	primaria	<3000	Nucleo	si	mt	24.8
8	Fortozo Campa Juan Luis	H	18	bgo	7.4	3	no	no	b	b	904	primaria	primaria	<3000	Nucleo	no	no	29.6
9	Gil Sereno Alejandro	H	19	cbta	8	1	no	si	b	b	958	sin estudios	sin estudios	<3000	Extensa	no	no	0
10	Gómez Ortega Carlos	H	19	ceplc	7.4	1	no	no	b	b	982	superior	media superior	3000-6000	Extensa	no	no	13.9
11	Hernández Morales Marco Antonio	H	18	bgo	8.6	1	no	no	i	b	970	primaria	primaria	<3000	Extensa	no	no	29.6
12	Herrera Hernández Ismael	H	19	bgo	7.2	1	no	si	b	b		no sabe	primaria	<3000	Extensa	no	no	
13	López Rongullo Flaviano	H	18	cobach	7	0	no	no	b	b	904	primaria	primaria	<3000	Extensa	no	no	24.8
14	Marin Rojano Carlos Gerardo	H	18	bgo	7.6	1	no	no	i	b	1042	secundaria	secundaria	<3000	Extensa	no	no	90
15	Mejía Méndez Oscar	H	17	cobach	7	3	no	si	b	b	1006	primaria	primaria	<3000	Extensa	no	no	29.6
16	Nájera Gil Mónica	M	18	conalep	8.4	0	no	si	i	b	910	media superior	media superior	<3000	Extensa	no	no	24.8
17	Orea Reyes Christian Uriel	H	17	bgo	7.6	1	no	si	b	b	976	secundaria	secundaria	<3000	Extensa	no	no	90
18	Paredes Quisehuaitl José Juan	H	19	bgo	8.5	0	no	si	b	b	964	sin estudios	primaria	<3000	Extensa	no	no	24.6
19	Pérez Landino Edith	M	18	bgo	9.2	0	no	no	b	b		primaria	primaria	<3000	Extensa	no	no	
20	Ramos Lázaro Luz María	M	19	bgo	8	1	no	si	b	b	934	primaria	primaria	<3000	Extensa	no	no	16.7
21	Rendón Galicia Jennifer	M	18	bgo	7.4	1	no	no	b	b	952	primaria	primaria	<3000	Extensa	no	no	19.5
22	Romero Degante Guadalupe	M	18	conalep	8.5	0	no	si	b	b		secundaria	secundaria	<3000	Extensa	no	no	
23	Rosas Torres Juan	H	18	cbtis	7.2	1	si	no	b	b	946	secundaria	secundaria	<3000	Nucleo	no	no	0
24	Sánchez Blanco Benjamín	H	18	cobach	7.4	1	no	si	b	b	904	no sabe	secundaria	<3000	Extensa	no	no	0
25	Sandoval Alvarez Efren	H	20	cobach	7.6	9	no	no	i	b	898	primaria	primaria	<3000	Extensa	si	mt	75
26	Tomás Vicencio Gabriela	M	20	bgo	7.5	1	no	si	b	b	916	primaria	primaria	<3000	Extensa	no	no	16.7
27	Vilchis Angel Iván	H	18	bgo	6.8	2	no	no	i	b	1006	primaria	primaria	<3000	Extensa	no	no	29.6
28	Zapata Aguilar Emanuel de Jesús	H	19	bgo	7.3	2	no	no	i	i	946	secundaria	secundaria	<3000	Nucleo	si	mt	0

Figura 3. Resultados de la prueba del modelo con datos reales

5 Resultados

La interfaz creada se encuentra disponible en el sitio Web de la UTIM [11] y cada tutor podrá hacer uso de ella.

En resumen, nuestra investigación mostró que los alumnos de la UTIM desertan por las siguientes tres causas principales:

- ✓ La *edad* es un factor importantísimo que tiene que ver con la madurez y perspectiva de futuro de los estudiantes,
- ✓ Los *ingresos familiares*, para aquellos alumnos cuya edad sea *menor o igual a 18 años*, puesto que a esta edad aún dependen de los ingresos familiares para el costo de su educación, y
- ✓ El *nivel de inglés*, para aquellos alumnos cuya edad sea mayor a 18 años.

Conclusiones

Como primera conclusión podemos afirmar que las técnicas de minería de datos que usamos proporcionan una manera que permite determinar aquellos alumnos que son candidatos a desertar. Existe la suficiente evidencia para afirmar que mediante la interfaz propuesta en esta investigación los tutores de nuestra institución podrán determinar este factor de riesgo de manera oportuna, para así dar seguimiento a

aquellos estudiantes vulnerables. Esta herramienta tiene la particularidad de que fue creada específicamente para los alumnos del Programa Educativo de TIC-SI, con datos históricos de éste y sólo puede utilizarse en la UTIM. La aplicación de este modelo en otros entornos no sería posible, sin embargo, el desarrollo de todo el proceso de descubrimiento del conocimiento y la aplicación de estas técnicas de minería de datos podría emularse de la misma manera.

Referencias

[1] FAYYAD, U. M., 1996: "Data Mining and Knowledge Discovery: Making Sense out of Data", IEEE Intelligent Systems, Vol. 11, No. 5, USA, ISSN: 0885-9000

[2] ANUIES, 2003, "El significado de la tutoría académica en estudiantes de primer ingreso a la licenciatura", Revista de la Educación Superior, Vol 3, No 127, México, ISSN: 01852760

[3] OCDE, 2006, "Higher education: quality, equity and efficiency", Obtenido el día 15 de Diciembre de 2009, desde la World Wide Web en el sitio http://www.oecd.org/site/0,3407,en_21571361_36507471_1_1_1_1_1,00.html

[4] BRITOS P., HOSSIAN A., 2005, "Minería de Datos", Nueva Librería, Argentina, ISBN: 9871104308

[5] KIMBALL, R, 2002, "The data warehouse toolkit: the complete guide to dimensional modeling", Wiley Computer Publishing, USA, ISBN: 780471200246

[6] CABENA, P., HADJINIAN, P., 1988, "Discovering Data Mining, From Concept to Implementation", Prentice Hall, USA, ISBN: 9780137439805

[7] MORALES, E., 2009, "Descubrimiento de conocimiento en bases de datos", Obtenido el día 11 de Julio de 2009, desde la World Wide Web en el sitio http://ccc.inaoep.mx/~emorales/Cursos/KDD/principal.html

[8] HAN J., KAMBER M., 2006, "Data mining: concepts and techniques", The Morgan Kaufmann Publishers, USA, ISBN: 1558609016

[9] HERNÁNDEZ J., FERRARI C., RAMÍREZ M., 2004. "Introducción a la minería de datos", Pearson Educación, España, ISBN: 9788420540917

[10] WITTEN I., FRANK E., 2005, "Data mining, practical machine learning, tools and learning", The Morgan Kaufmann Publishers, USA, ISBN: 0120884070

[11] VALERO S., SALVADOR A., GARCÍA M., 2009, "Modelo de Predicción", Disponible en la World Wide Web en http://www.utim.edu.mx/mineria