Análisis Comparativo de Algoritmos de Minería de Datos para Predecir la Deserción Escolar

Maricela Quintana López, Juan Carlos Trinidad Pérez, Saturnino Job Morales Escobar, Víctor M. Landassuri Moreno

Centro Universitario UAEM Valle de México
{mquintanal, sjmoralese, vlandassurim}@uaemex.mx,
jtrinidadp359@alumno.uaemex.mx

Resumen. Un tutor académico, debe dar seguimiento personalizado al alumno para evitar que deserte por cuestiones académicas, por ejemplo, materias que se le dificulten. Sin embargo, la información del seguimiento es bastante, por lo que debe procesarse para extraer patrones que permitan tomar decisiones oportunamente. Este artículo presenta la comparación de algoritmos de clasificación para la predicción de la deserción escolar del Centro Universitario UAEM Valle de México en la carrera de Ingeniería en Sistemas y Comunicaciones. El objetivo fue determinar el mejor algoritmo basándose en la precisión de la clasificación, así como en la utilidad en la información provista al tutor. De los experimentos concluimos que los mejores algoritmos son Naïve Bayes Tree, que tiene el menor error, y J48 que provee información útil para el tutor al indicar las materias que debe cuidar, permitiéndole crear estrategias que apoyen el desempeño académico del alumno en las mismas.

Palabras clave: Minería de datos, clasificación, árboles de decisión, naïve bayes.

1 Introducción

La deserción escolar en las instituciones de nivel superior, es un problema que actualmente las autoridades administrativas enfrentan. Diversos estudios realizados llegan a la conclusión de que el problema existe tanto a nivel local como a nivel global y este continuará si no se atiende de manera adecuada.

Para cada país o estado existen estadísticas que describen el nivel de deserción estudiantil en educación superior, permitiéndonos detectar la situación que se vive para cada una de las carreras profesionales de las instituciones públicas y privadas [18-20]. Sin embargo, es posible utilizar las tecnologías de la información para detectar situaciones de riesgo que pueden llevar a un alumno a abandonar sus estudios. Ésta fue la motivación principal por la cual en la Universidad Autónoma del Estado de México, se diseñó el sistema SITA (Sistema Inteligente para la Tutoría Académica), que permite a los tutores académicos monitorear el desempeño académico y consultar información socioeconómica del alumno [16-17].

Advances in Computer Science and Engineering
Research in Computing Science 67, 2013 pp. 13–23
Paper Recived 30-08-2013 and Acepted 21-10-2013

A pesar de este gran apoyo, determinar si un alumno desertará o no, requiere de un análisis particular para cada trayectoria y situación, labor que debe realizar el tutor. Por ello, en este trabajo nos enfocamos en generar un modelo clasificador cuyo nivel de fiabilidad y precisión sea aceptable para determinar la posibilidad de que un alumno interrumpa sus estudios. Con la gran ventaja de que la aplicación del modelo sea realizada de manera automática.

Para realizar este trabajo, se tomará como base las calificaciones obtenidas por 193 alumnos de la carrera de Ingeniería en Sistemas y Comunicaciones durante el primer año de estudios, específicamente de las generaciones 2008 (71 alumnos), 2009 (63 alumnos) y 2010 (59 alumnos).

En la figura 1, se muestran las diferentes fases de la metodología utilizada en este trabajo, la cual está basada en el modelo KDD (Knowledge Discovery in Databases) [9,11].

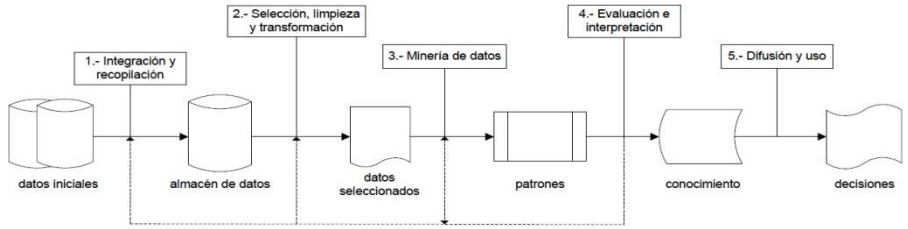

Fig. 1. Fases del proceso KDD

Con esto en mente, en la sección 2 se mostrará la preparación de los datos, seguida de la sección 3 en donde se presentarán las técnicas de minería de datos empleadas en esta investigación. Posteriormente, en la sección 4 se mostrarán los experimentos y resultados obtenidos. Finalmente, en las secciones 5 y 6 se presentarán las conclusiones y trabajo futuro respectivamente.

2 Preparación de los datos

En este trabajo, se utilizaron las trayectorias académicas de todos los alumnos de la carrera de Ingeniería en Sistemas y Comunicaciones, de las generaciones 2008, 2009 y 2010. La información obtenida de las bases de datos de control escolar fue filtrada para utilizar únicamente las calificaciones de las unidades de aprendizaje del primer año de estudios que cursaron los 193 alumnos. En la tabla 1, se muestran los atributos y los posibles valores numéricos y nominales de los mismos. El atributo Número de Cuenta se utiliza únicamente para identificar los registros más no para generar los patrones, mientras que el atributo desertó es la clase a aprender.

Es necesario aclarar que las calificaciones consideradas son únicamente las primeras que obtuvo el alumno en las unidades de aprendizaje, no consideramos las notas obtenidas en evaluaciones extraordinarias. También, fueron eliminados de la muestra

los registros de alumnos que únicamente cursaron el primer semestre, ya que el trabajo considera lo que ocurre en el primer año.

Tabla 1. Datos utilizados con su respectivo tipo de variable

Atributos	Valores Numéricos	Valores Nominales
Número de Cuenta	Matrícula institucional	Matrícula institucional
Introducción a la Computación	0-100	A: Aprobado R: Reprobado NP: No Presentó
Álgebra y Geometría Analítica		
Administración		
Estática y Dinámica		
Introducción a la Ingeniería		
Técnicas de Comunicación		
Álgebra Lineal		
Calculo Diferencial e Integral		
Fundamentos de Programación		
Arquitectura de computadores		
Química		
Desertó	si, no	si, no

Debido a la naturaleza de los algoritmos a aplicar, se transformaron los datos para tenerlos de forma numérica y de forma nominal. Para la versión numérica, las calificaciones NP fueron transformadas a cero. En la tabla 2, se muestra un fragmento de cómo quedaron los datos.

Tabla 2. Muestra de datos a considerar en el análisis con datos numéricos

	Primer Semestre				Segundo Semestre			Clase
Cuenta	**Introducción a la Computación**	**Técnicas de Comunicación**	**Administración**		**Álgebra Lineal**	**Química**	**Arquitectura de Computadores**	**Desertó**
449905	65	47	32		70	81	48	Si
540214	92	100	70	…	92	82	88	No
726037	92	71	100		83	87	90	No
823719	68	100	70		71	67	60	No
823753	73	90	70		60	87	82	Si

Para la versión nominal, las calificaciones mayores o iguales a 60 tienen un valor de A (aprobado), las menores a 60, tienen un valor de R (reprobado) y las evaluaciones no presentadas tienen un valor de NP. En la tabla 3, se muestra un conjunto de datos con la transformación mencionada.

Tabla 3. Muestra de datos a considerar en el análisis con datos nominales

	Primer Semestre				Segundo Semestre			Clase
Cuenta	**Introducción a la Computación**	**Técnicas de Comunicación**	**Administración**		**Álgebra Lineal**	**Química**	**Arquitectura de Computadores**	**Desertó**
449905	A	R	R		A	A	R	Si
540214	A	A	A	…	A	A	A	No
726037	A	A	A		A	A	A	No
823719	A	A	A		A	A	A	No
823753	A	A	A		A	A	A	Si

3 Minería de datos

Existen varios trabajos desarrollados para predecir el desempeño académico de un alumno empleando técnicas de minería de datos. Entre las técnicas más utilizadas están los árboles de decisión y las técnicas bayesianas [1-8].

Un árbol de decisión es un conjunto de condiciones organizadas de forma jerárquica para clasificar una serie de atributos y predecir el valor de la clase. Los árboles están formados por nodos y ramas donde cada nodo representa una condición sobre algún atributo y cada rama corresponde a un posible valor para ese atributo. Pueden manejar valores de tipo nominal y/o de tipo numérico [10, 12]. Los algoritmos de árboles de decisión más utilizados son ID3, C4.5 (J48) y Naïve Bayes Tree [1-8]. Cabe aclarar que el ID3 no trabaja con datos numéricos, únicamente con categóricos.

Por otra parte, los algoritmos Naïve Bayes son un grupo de clasificadores estadísticos que al ser aplicados a una instancia nueva, dan como resultado la probabilidades de que dicha instancia pertenezca a una clase determinada [11, 13]. Algunos de los trabajos realizados utilizan principalmente el algoritmo de Naïve Bayes y las redes Bayesianas [2, 3, 10, 15]. Los algoritmos basados en Bayes funcionan de manera correcta con datos del tipo nominal y/o numérico y son independientes.

Los árboles de decisión y los algoritmos Bayesianos han demostrado tener un alto índice de precisión cuando de predicción se trata, de hecho han sido incluidos dentro de los diez mejores algoritmos de minería de Datos [14]. Por ello en este trabajo se realiza un análisis comparativo entre algoritmos de árboles de decisión y Bayesianos, particularmente, ID3 (solo para valores nominales), J48 (C4.5), Naïve Bayes Tree, Naïve Bayes, y redes Bayesianas, los cuales se encuentran disponibles en el software WEKA de la universidad de Waikato, de Nueva Zelanda [21].

4 Experimentos y resultados obtenidos

Las pruebas se realizaron en dos grandes bloques: uno para los datos de tipo nominal y otro para los datos de tipo numérico. En ambos casos, se aplicaron los algoritmos elegidos para conducir 3 experimentos que varían en cuanto a los datos empleados en los mismos y que se explican a continuación. Todos los modelos fueron evaluados

utilizando la validación cruzada de 10 pliegues, la cual divide el conjunto de datos en diez partes y utiliza nueve partes para entrenamiento y una para prueba. El proceso es repetido diez veces [10-11].

En el primer experimento realizado se aplicaron los algoritmos elegidos a los datos de cada generación por separado para construir su modelo. Las tablas 4 y 5 muestran los resultados obtenidos por cada algoritmo, de la siguiente forma: porcentaje de precisión, instancias clasificadas (correctamente / incorrectamente).

Tabla 4. Resultados utilizando datos nominales

Algoritmo	**Porcentaje de precisión por generación usando datos nominales**			**Promedio**
	2008	**2009**	**2010**	
ID3	95.7746 (68/71)	85.4839 (53/63)	98.3051 (58/59)	93.19
J48	88.7324 (63/71)	88.7097 (55/63)	98.3051 (58/59)	91.92
Naïve Bayes Tree	92.9577 (66/71)	83.871 (52/63)	96.6102 (57/59)	91.15
NaïveBayes	92.9577 (66/71)	87.9068 (54/63)	96.6102 (57/59)	92.49
BayesNet	91.5493 (65/71)	87.9068 (54/63)	96.6102 (57/59)	92.02

Podemos observar que para los datos de tipo nominal, si consideramos las generaciones 2008 y 2010, el mejor algoritmo resulta ser ID3. Pero si consideramos 2009 y 2010, el mejor resulta ser el algoritmo J48. Al obtener el promedio de las 3 generaciones podemos notar que el mejor es el ID3 para valores de tipo nominal. Por otro lado, cuando se trata de datos numéricos el mejor algoritmo en las 3 generaciones es el Naïve Bayes Tree. Al comparar los resultados de los categóricos con los numéricos, resulta mejor el Naive Bayes Tree, lo cual es comprensible si consideramos que los datos son de naturaleza numérica y al ser transformados para trabajarlos como categóricos, se pierde la precisión.

Tabla 5. Resultados obtenidos utilizando datos numéricos

Algoritmo	**Porcentaje de precisión por generación usando datos numéricos**			**Promedio**
	2008	**2009**	**2010**	
ID3	NA	NA	NA	NA
J48	88.7324 (63/71)	90.4762 (57/63)	98.3051 (58/59)	92.50457
Naïve Bayes Tree	92.9577 (66/71)	92.0635 (58/63)	98.3051 (58/59)	94.4421
NaïveBayes	85.9155 (61/71)	74.6032 (47/63)	98.3051 (58/59)	86.2746
BayesNet	84.507 (60/71)	92.0635 (58/63)	98.3051 (58/59)	91.6252

Como una muestra de los modelos generados, en las figuras 2 y 3 se presentan los árboles de decisión para las generaciones 2008 y 2009 utilizando los algoritmos de ID3 y J48 respectivamente. En ellas podemos observar que la materia que provee mayor información para la separación de las clases es Arquitectura de Computadores.

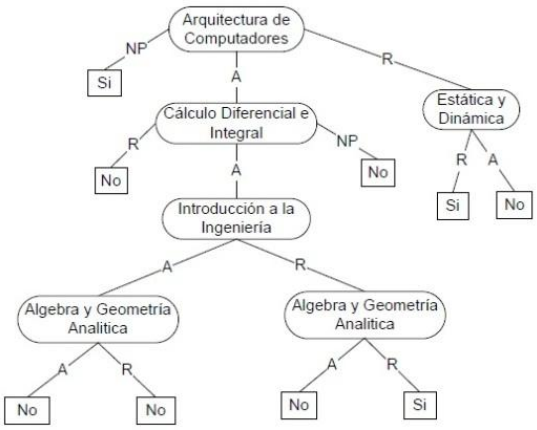

Fig. 2. Árbol para la generación 2008 usando ID3.

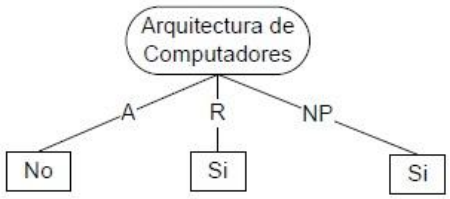

Fig. 3. Árbol para la generación 2009 usando J48(C4.5).

Como un segundo experimento se aplicaron los algoritmos elegidos a todo el conjunto de datos, es decir, generaciones 2008-2010 y los resultados referentes a la precisión obtenida se muestran en la tabla 6.

Tabla 6. Resultados obtenidos al combinar las tres generaciones, 2008-2010

Algoritmo	% de precisión	
	Variables de tipo nominal	Variables de tipo numérico
ID3	86.4583	NA
J48	90.625	88.7755
Naïve Bayes Tree	91.667	90.3061
NaïveBayes	89.5833	82.6530
BayesNet	89.0625	87.7551

Podemos observar que en este caso, el algoritmo con mayor precisión es el Naïve Bayes Tree tanto para variables nominales como numéricas. El árbol obtenido usando ID3 en este conjunto de datos se muestra en la figura 4, en él podemos observar que la unidad de aprendizaje Arquitectura de computadores sigue siendo la que se encuentra en el nodo raíz del árbol.

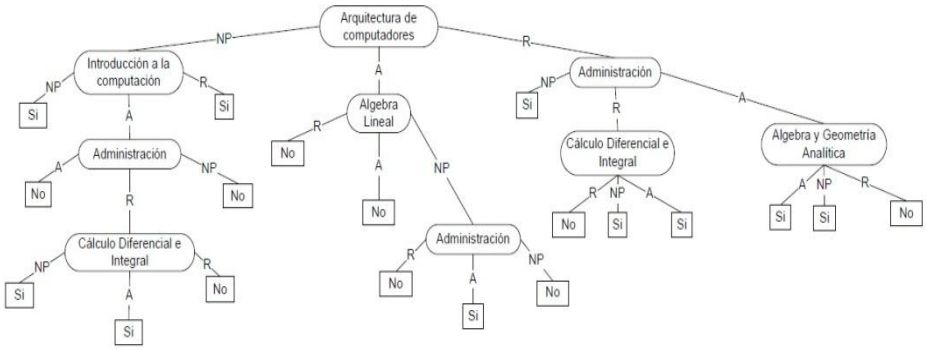

Fig. 4. Árbol para las generaciones 2008-2010 usando ID3.

En cuanto al algoritmo Naïve Bayes Tree para valores numéricos, el árbol generado es de un nivel y se muestra en la figura 5. Podemos notar que la unidad de aprendizaje en la raíz es la de Arquitectura de computadores, y las hojas son los clasificadores bayesianos.

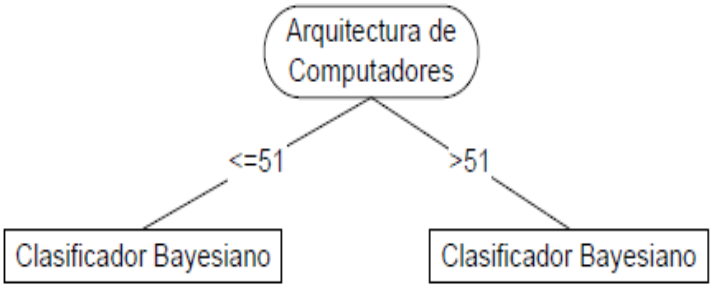

Fig. 5. Árbol para las generaciones 2008-2010 usando Naïve Bayes Tree

De acuerdo a los resultados obtenidos, fue detectado que la calificación nominal NP estaba provocando ruido en el sistema, por lo que se realizó un tercer experimento en el que se transformó el NP a reprobado R, pues un alumno con calificación NP tendrá que volver a cursar forzosamente la unidad de aprendizaje, como si estuviese reprobado. Los resultados obtenidos al realizar este cambio se muestran en la tabla 7.

Tabla 7. Resultados obtenidos al combinar las tres generaciones, 2008-2010, sin usar NP

Algoritmo	% de precisión en las variables de tipo nominal
ID3	88.5417
J48	88.0208
Naïve Bayes Tree	90.625
NaïveBayes	89.0625
BayesNet	89.0625

Los árboles generados por Naïve Bayes Tree y J48, se muestran en las figuras 6 y 7. En ellas podemos apreciar de manera más clara, qué grupos de materias llevan, en conjunto, al abandono de los estudios por parte del alumno.

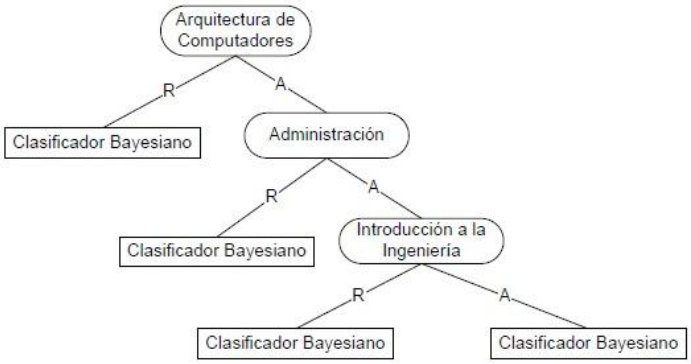

Fig. 6. Árbol generado por Naïve Bayes Tree sin usar NP

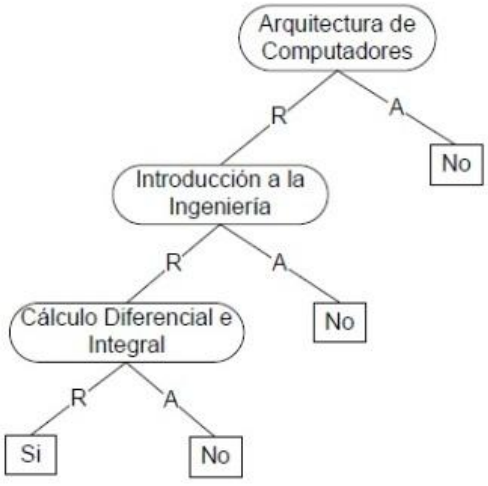

Fig. 7. Árbol generado por J48 sin usar NP

5 Conclusiones y trabajo futuro

En este trabajo fue presentado un análisis comparativo entre varias técnicas de minería de datos para determinar si un alumno es propenso a desertar de sus estudios en la carrera de Ingeniería en sistemas y Comunicaciones del Centro Universitario UAEM Valle de México. Los resultados obtenidos muestran que es posible obtener un modelo de predicción de la deserción escolar utilizando las calificaciones obtenidas por los alumnos en el primer año de sus estudios.

En los experimentos realizados con los datos de las generaciones, podemos concluir que los mejores algoritmos son Naíve Bayes Tree y J48. El algoritmo de Naïve Bayes Tree es, desde el punto de vista cuantitativo, el que tiene el menor error al momento de clasificar. Sin embargo, desde el punto de vista cualitativo, el árbol de inducción generado por el algoritmo J48 provee información que puede ser de utilidad para el tutor académico, pues le indica las unidades de aprendizaje en las que debe poner atención y planear estrategias que apoyen el desempeño académico del alumno.

Como trabajo futuro, se utilizarán los datos de más generaciones para hacer un estudio a detalle del comportamiento académico histórico que existe en la carrera de Ingeniería en Sistemas y Comunicaciones. También podrían incluirse en el análisis otros algoritmos de clasificación como CART, AdaBoost, SVM entre otros.

Referencias

1. Vera, C. M., Morales, C. R., & Soto, S. V.: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem, 109. (2012).
2. Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. A Comparative Analysis of Techniques for Predicting Academic Performance. Paper presented at the 37th ASEE/IEEE Frontiers in Education Conference. October 10 – 13, Milwaukee, WI. (2007).
3. García, E. P. I., & Mora, P. M.: Model Prediction of Academic Performance for First Year Students. In Artificial Intelligence (MICAI), 2011 10th Mexican International Conference on (pp. 169-174). IEEE. (2011).
4. Vasile Paul Bresfelean: Data Mining Applications in Higher Education and Academic Intelligence Management, Theory and Novel Applications of Machine Learning, Meng Joo Er and Yi Zhou (Ed.), ISBN: 978-953-7619-55-4, InTech, DOI: 10.5772/6684. (2009).
5. Alcover, R., Benlloch,J., Blesa P., Calduch, M. A., Celma, M., Ferri, C.: Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. XIII Jornadas de Enseñanza universitaria de la Informática, Teruel, España. (2007).
6. Orea, S. V., Vargas, A. S., & Alonso, M. G.: Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Ene, 779(73), 33. (2005).
7. Spositto, O. M., Etcheverry, M. E., Ryckeboer, H. L., & Bossero, J. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. Memorias de la 9ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2010. Orlando, Florida, USA, (2010).

8. Pereira, R. T. Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. Memorias de la 8ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2009. Orlando, Florida, USA, (2009).
9. Moine, J. M., Haedo, A., & Gordillo, S.: Estudio comparativo de metodologías para minería de datos. In XIII Workshop de Investigadores en Ciencias de la Computación. (2011).
10. Sharma M., Mavani M.: Development of predictive model in education system: using naïve bayes classifier. International conference and workshop on emerging trends in technology-ICWET, (2011).
11. Orallo, J. H., Quintana, M. J. R., & Ramírez, C. F.: Introducción a la Minería de Datos. Pearson Prentice Hall. (2004).
12. Han, J., Kamber, M., & Pei, J.: Data mining: concepts and techniques. Morgan kaufmann. (2006).
13. Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D.: Artificial intelligence: a modern approach (Vol. 74). Englewood Cliffs: Prentice hall, (1995)
14. Wu, X., & Kumar, V. (Eds.). (2010). The top ten algorithms in data mining. CRC Press.
15. Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L.: A case study of knowledge discovery on academic achievement, student desertion and student retention. In Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on (pp. 150-154). IEEE, (2004).
16. Morales S. J., Rodríguez I, García S.: Propuesta de un sistema de información para la Tutoría Académica en una institución de educación superior. III Congreso Internacional de Innovación Educativa. Xalapa, Veracruz. México, (2008).
17. Sistema Inteligente para la Tutoría Académica, https://www.sita.uaemex.mx/tutoria/
18. Dirección General de Educación Superior Universitaria, http://www.dgesu.ses.sep.gob.mx/Principal/subdirecciones/indicadores/desercion.aspx
19. Universidad Autónoma del Estado de México, http://www.uaemex.mx/planeacion/NumEstadis.html.
20. Publicaciones ANUIES, http://publicaciones.anuies.mx/
21. WEKA 3: Data Mining software in java, http://www.cs.waikato.ac.nz/ml/weka/