Artículo **Revista de Sistemas y Gestión Educativa**
Marzo 2015 Vol.2 No.2 206-213

Análisis de los resultados del EXANI-II en el Estado de Aguascalientes mediante técnicas de minería de datos

LUNA-RAMÍREZ, Enrique*†`, CORREA-VILLALÓN, Christian`, VELARDE-MARTÍNEZ, Apolinar` y HERNÁNDEZ-CHESSANI, David``

`*Instituto Tecnológico El Llano Aguascalientes. Km. 18 Carr. Ags.-S.L.P. El Llano Aguascalientes. C.P. 20330.*
``*Universidad Tecnológica de Aguascalientes. Boulevard Juan Pablo II 1302 Aguascalientes, Ags. C.P. 20206.*

Recibido 9 de Enero, 2015; Aceptado 5 de Marzo, 2015

Resumen

Las técnicas de minería de datos permiten obtener el conocimiento oculto en los grandes volúmenes de datos generados en cualquier contexto, particularmente en el contexto educativo. Con la ayuda de este tipo de técnicas y de herramientas especializadas se está llevando a cabo un análisis de las bases de datos del EXANI-II del Estado de Aguascalientes correspondientes al año 2013, cuyo propósito principal es la identificación de factores que impactan de manera negativa el desempeño académico de los estudiantes de nivel medio superior, así como la definición de estrategias para fortalecer los aspectos débiles que sean identificados en dicho desempeño. Los modelos generados como parte fundamental de este estudio serán validados en un estudio posterior utilizando los datos del año 2014, de manera que éstos, los modelos, cuenten con un alto índice de confiabilidad en su utilización. Un análisis preliminar de los datos del año 2013 ha sugerido utilizar principalmente las técnicas de *clasificación* (arboles de decisión) y *clustering* (agrupación por sectores) en este estudio.

Minería de datos educativa, EXANI-II

Abstract

Data mining techniques allow extracting the hidden knowledge in big data sets generated in any field, particularly in the educational field. With the help of this kind of techniques and specialized tools, it is being carried out an analysis of the EXANI-II data bases of the Aguascalientes State (México) corresponding to the 2013 year, whose main purpose is the identification of the factors that impact negatively the academic performance of senior high students, as well as the definition of strategies to reinforce the weak aspects identified in this performance. The models generated as fundamental part of this study will be validated in a subsequent study by using the data corresponding to the 2014 year, in such a way that the generated models have a high level of confidence at the moment of being used. A preliminary data analysis has suggested using mainly the techniques of *classification* (decision trees) and *clustering* (grouping by sector) in this study.

Educational data mining, EXANI-II

Citación: LUNA-RAMÍREZ, Enrique, CORREA-VILLALÓN, Christian, VELARDE-MARTÍNEZ, Apolinar y HERNÁNDEZ-CHESSANI, David**.** Análisis de los resultados del EXANI-II en el Estado de Aguascalientes mediante técnicas de minería de dato. Revista de Sistemas y Gestíon Educativa 2015, 2-2:206-213

* Correspondencia al Autor (Correo Electrónico: elunaram@hotmail.com)
†Investigador contribuyendo como primer autor

Introducción

Teniendo en cuenta el auge de la *minería de datos* como una alternativa eficaz para el análisis de datos y la imperante necesidad de mejorar la calidad educativa en nuestro país, particularmente en el Estado de Aguascalientes, la aplicación de técnicas de minería de datos en los datos generados de exámenes coordinados por el Centro Nacional de Evaluación para la Educación Superior (CENEVAL, 2015) representa un área de oportunidad importante para detectar y corregir deficiencias en la población estudiantil en diferentes niveles educativos. Es importante señalar que este tipo de estudios quedan enmarcados en la denominada *minería de datos educativa*, campo que ha venido cobrando interés en la comunidad científica.

El estudio del desempeño académico de los estudiantes (de diferentes niveles) no es un ejercicio nuevo, no obstante, la forma de analizar los datos generados a partir de diversas evaluaciones ha mejorado en los últimos años con la incorporación de técnicas de análisis de datos novedosas como es el caso de la *minería de datos*, cuyas técnicas pueden ser utilizadas para predecir el desempeño académico de los estudiantes, entre otras variables que pudieran contribuir al diseño de nuevas estrategias para mejorar la calidad educativa.

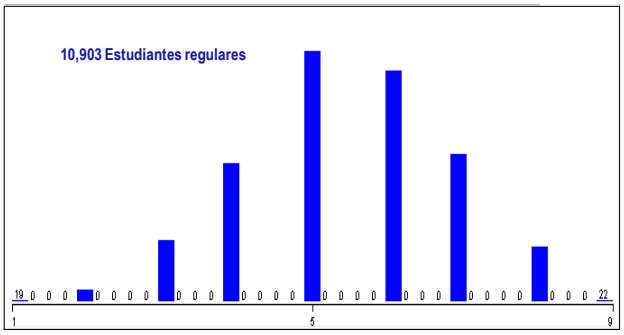

Figura 1 Distribución de promedios generales de estudiantes regulares del EXANI-II en el año 2013 en el Estado de Aguascalientes

En la figura 1 se presenta, sin detallar variables, la distribución de los promedios generales (ajustados a la escala de 0 a 10) de los 10,903 estudiantes regulares que sustentaron el EXANI-II en el Estado de Aguascalientes en el año 2013. Como es natural, los resultados tienden a seguir una distribución Normal.

Existe otro grupo complementario de 5,849 estudiantes, quienes en su momento no sustentaron por diversas razones el EXANI-II, razón por la cual el Instituto de Educación de Aguascalientes ofreció otro periodo para que estos estudiantes rezagados sustentaran dicho examen. En la figura 2 se presentan los resultados generados.

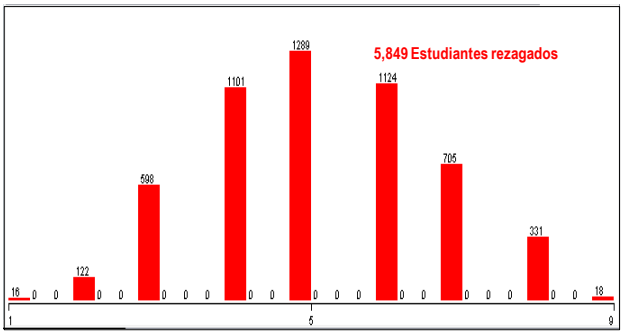

Figura 2 Distribución de promedios generales de estudiantes rezagados del EXANI-II en el año 2013 en el Estado de Aguascalientes

Como se puede observar, las dos gráficas anteriores presentan cierta similitud, por lo que cabe la posibilidad de que al momento de generar los modelos mediante técnicas minería de datos, éstos puedan ser generados con un conjunto de datos y validados con el otro conjunto.

Revisión de literatura

Algunos de los trabajos más importantes realizados en torno al tópico de la *minería de datos educativa* se describen a continuación:

ISSN 2410-3977

Bresfelean (2007) realizó un estudio para predecir la elección de carrera profesional de estudiantes de diferentes especialidades, para lo cual desarrolló un conjunto de arboles de decisión basados en el algoritmo *J48* de WEKA, argumentando que en general este algoritmo arroja mejores resultados que otros algoritmos tales como el algoritmo ID3.

Cheewaprakobkit (2013) realizó un estudio para identificar "estudiantes débiles", de manera que el desempeño académico de tales estudiantes pueda ser mejorado. En su estudio, los autores utilizan dos algoritmos: *La Red Neuronal* y el *C4.5 Tree*. El desarrollo de su trabajo consistió de tres etapas principales: preprocesamiento de datos, filtrado de atributos y generación de reglas de clasificación. Concluyen su estudio identificando a la técnica de árboles de decisión como la técnica más eficiente para clasificar los datos disponibles.

Kumar y Vijayalakshmi (2011) proponen un enfoque para predecir el desempeño de estudiantes en ciertos tipos de evaluación. Utilizan el algoritmo *C4.5* (J48 en WEKA) para llevar a cabo su análisis predictivo. En la colección de datos, hacen una ligera adecuación en la definición de los valores nominales y, su vez, los valores enteros son transformados en valores nominales. Los datos en su conjunto son almacenados en formato .CSV y posteriormente son llevados al formato .ARFF de WEKA. La implementación de las reglas en los arboles de decisión generados son extraídas dividiendo los datos en dos grupos.

Pal y Pal (2013) presentan un enfoque de clasificación para predecir la colocación de estudiantes. Este enfoque provee las relaciones entre registros académicos y la colocación de estudiantes. En este análisis, se emplean diversos algoritmos de clasificación utilizando herramientas de minería de datos como WEKA y el proceso de entrenamiento utiliza un conjunto de atributos predefinidos.

Los algoritmos de clasificación más ampliamente utilizados en este trabajo son *Naïve Bayes*, *Multilayer Perceptron* y *C4.5 Tree*, siendo éste último el más popular debido a sus características agregadas como la supervisión de valores faltantes, la categorización de atributos continuos, la poda de árboles de decisión, etc.

Ramanathan *et al.* (2013) llevaron a cabo un estudio sobre el desempeño académico de estudiantes utilizando el algoritmo ID3 modificado, para lo cual debieron corregir algunos defectos del algoritmo ID3 (utilizado para generar arboles de decisión). Este algoritmo modificado es denominado WID3 (ponderado). Los autores concluyen afirmando que este algoritmo modificado resultó más eficiente que los algoritmos J48 y Naïve Bayes.

Marco teórico

El EXANI-II proporciona información integral sobre quiénes son los aspirantes que cuentan con mayores posibilidades de éxito en los estudios de nivel superior y cuál es su nivel de desempeño en áreas fundamentales para el inicio de los estudios superiores o de técnico superior universitario. Este examen integra dos pruebas:

EXANI-II Admisión, que explora competencias genéricas predictivas en las áreas de pensamiento matemático, pensamiento analítico, estructura de la lengua y comprensión lectora. Su propósito es establecer el nivel de potencialidad de un individuo para lograr nuevos aprendizajes, por lo que todo sustentante debe responderlo. Ofrece a las instituciones información útil para la toma de decisiones sobre la admisión de los aspirantes.

EXANI-II Diagnóstico, que mide el nivel de la población sustentante en el manejo de competencias disciplinares, alineadas con la Reforma Integral de Educación Media Superior.

ISSN 2410-3977

Dado su carácter diagnóstico, la institución usuaria tiene la prerrogativa de incluir o no esta prueba en su proceso de aplicación.

Como punto de partida de este estudio, ha sido necesario determinar cuáles de las 98 variables incluidas en el catálogo del EXANI-II son relevantes para los propósitos del mismo, para lo cual se realizaron diversas pruebas de pertinencia, habiéndose definido las siguientes variables:

Variable	Descripción	Valores
ICNE	Calificación en índice CENEVAL del examen de selección	700-1300
PCNE	Calificación en porcentaje de aciertos del examen de selección	0%-100%
PRLM	Calificación de razonamiento lógico matemático en porcentaje de aciertos	0%-100%
PMAT	Calificación de matemáticas (selección) en porcentaje de aciertos	0%-100%
PRV	Calificación de razonamiento verbal en porcentaje de aciertos	0%-100%
PESP	Calificación de español en porcentaje de aciertos	0%-100%
PTIC	Calificación de tecnologías de información y comunicación en porcentaje de aciertos	0%-100%

Metodología

Para el desarrollo de este estudio, se ha seguido la metodología propia de la construcción de un Data Warehouse, a decir, la extracción, transformación y carga de los datos en un repositorio único (ETL, por sus siglas en inglés), y su posterior explotación mediante herramientas especializadas, aunque a una escala menor, dado que en un Data Warehouse se integran datos de diferentes contextos.

La primera acción realizada fue el análisis de las bases de datos del EXANI-II, operando el proceso ETL para seleccionar datos útiles y su posterior limpieza y transformación al formato .arff de WEKA con la finalidad de generar vistas minables que permitan generar modelos estadísticamente confiables.

Con base en la literatura, para llevar a cabo la tarea de minar los datos, se consideró utilizar en principio algoritmos reconocidos como efectivos en diversas situaciones tales como el *J48* de WEKA para clasificar los datos mediante arboles de decisión y el *K-means* para agrupar los datos de manera que se observen patrones claros en los grupos.

Los modelos generados fueron validados con una parte del conjunto original de datos, aunque está pendiente validarlos con los datos del EXANI-II del año 2014. No obstante, se extrajeron resultados (reglas) preliminares, algunas de las cuales den una idea clara del desempeño académico por sectores.

Este proyecto concluirá con la definición de estrategias y acciones para corregir deficiencias académicas que hayan sido detectadas, esto con el apoyo de expertos del Instituto de Educación de Aguascalientes.

Análisis exploratorio de datos

Un análisis estadístico previo de los datos sobre las variables de interés permitió tener una idea más clara de su comportamiento, al observarse una complementariedad natural entre los resultados de estudiantes regulares y rezagados.

ISSN 2410-3977

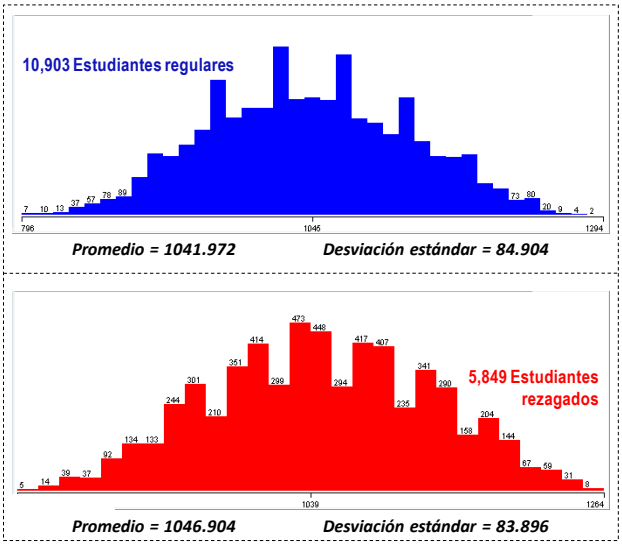

Figura 3 Distribución de calificaciones en índice CENEVAL (icne: 700-1300)

Las gráficas de la figura 3 sugieren fuertemente unir ambos conjuntos de datos con la finalidad de tener un conjunto de datos más "normalizado" y, por ende, un conjunto de datos más adecuado para generar modelos estadísticamente confiables y útiles. En la figura 4 se muestra el resultado de dicha unión, pudiéndose constatar que efectivamente se trata de un conjunto de datos más "normalizado".

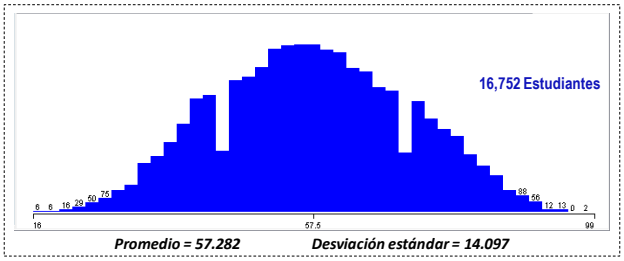

Figura 4 Distribución de calificaciones CENEVAL en porcentaje del total de estudiantes (pcne: 0%-100%)

Además de la mejora que se obtuvo en la distribución de los resultados globales al considerar el total de los estudiantes que sustentaron el EXANI-II, también se logró una mejora por variables específicas, situación mostrada en la figura 5.

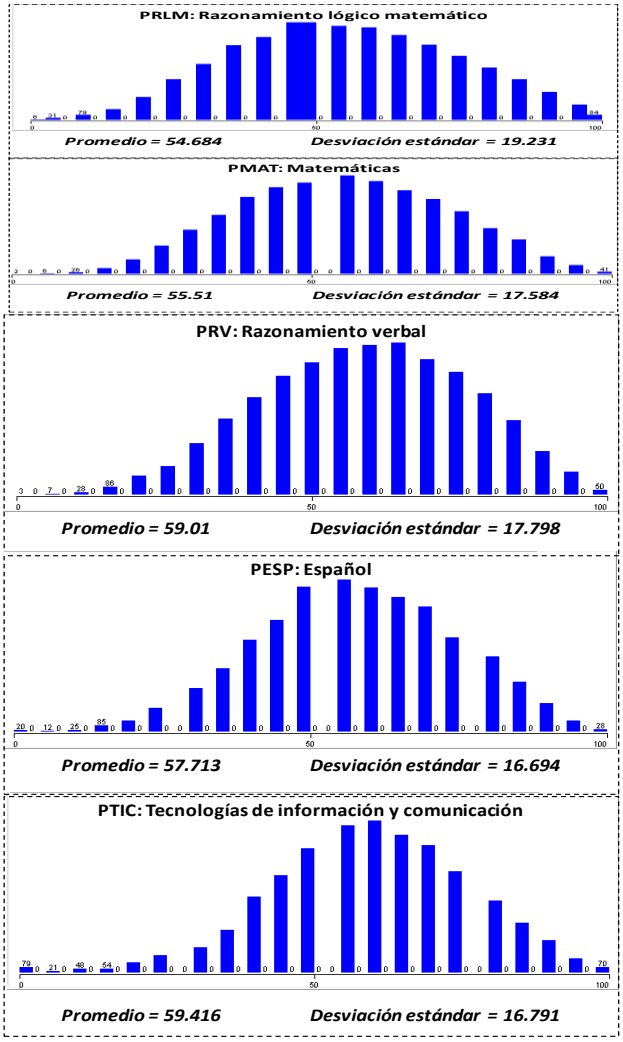

Figura 5 Distribución de calificaciones en porcentaje por áreas

En las cinco gráficas de la figura 5 se puede observar un comportamiento con tendencia normal de los datos.

Generación de modelos preliminares

Con base en el análisis exploratorio descrito en la sección anterior, se procedió a la generación de modelos utilizando WEKA, que hoy en día es una de las herramientas más populares para minar datos (The University of Waikato, 2015).

ISSN 2410-3977

En las figuras 6 y 7 se muestran los modelos obtenidos al aplicar la técnica de clasificación (árboles de decisión) utilizando el algoritmo *J48* y *RandomTree*, respectivamente.

En el primer modelo (figura 6), se pueden observar reglas potencialmente interesantes, pero con un alto porcentaje de clasificación mal realizada, concretamente más del 25% de los datos están mal clasificados, lo que sugiere utilizar un algoritmo diferente al *J48*.

```
Size of the tree :      820


Time taken to build model: 0.18 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        3949               74.8484 %
Incorrectly Classified Instances      1327               25.1516 %
Kappa statistic                          0.4958
Mean absolute error                      0.3198
Root mean squared error                  0.3999
Relative absolute error                 64.4449 %
Root relative squared error             80.2777 %
Total Number of Instances             5276
Ignored Class Unknown Instances                573

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
                 0.758     0.259      0.71      0.758     0.733
                 0.741     0.242      0.785     0.741     0.762
Weighted Avg.    0.748     0.25       0.751     0.748     0.749

=== Confusion Matrix ===

    a    b   <-- classified as
 1825  583 |    a = H
  744 2124 |    b = M


=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Resultados Rezagados 2013_9var-weka.filters.
Instances:    5849
Attributes:   9
              sexo
              nom_proc
              pcne
              prlm
              pmat
              prv
              pesp
              ptic
              pmei
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

nom_proc = PREPARATORIA ABIERTA: H (91.0/39.0)
nom_proc = COLEGIO DE BACHILLERES DEL ESTADO DE ZACATECAS
|   prlm <= 35
|   |   ptic <= 60
|   |   |   pesp <= 30
|   |   |   |   pmei <= 25: H (3.0)
|   |   |   |   pmei > 25
|   |   |   |   |   pmei <= 35: M (10.0/1.0)
|   |   |   |   |   pmei > 35
|   |   |   |   |   |   pesp <= 20: M (2.0)
```

Figura 6 Técnica de clasificación (*Classifying*) utilizando el algoritmo *J48* de WEKA

```
Size of the tree : 4080

Time taken to build model: 0.11 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        5276               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             5276
Ignored Class Unknown Instances                573

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
                 1         0          1         1         1
                 1         0          1         1         1
Weighted Avg.    1         0          1         1         1

=== Confusion Matrix ===

    a    b   <-- classified as
 2408    0 |    a = H
    0 2868 |    b = M

=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.
Relation:     Resultados Rezagados 2013_9var-weka.filters.
Instances:    5849
Attributes:   9
              sexo
              nom_proc
              pcne
              prlm
              pmat
              prv
              pesp
              ptic
              pmei
Test mode:    evaluate on training data

=== Classifier model (full training set) ===


RandomTree
==========

ptic < 57.5
|   nom_proc = PREPARATORIA ABIERTA
|   |   pmat < 67.5
|   |   |   prv < 77.5
|   |   |   |   ptic < 20 : H (2/0)
|   |   |   |   ptic >= 20
|   |   |   |   |   pcne < 49.5
|   |   |   |   |   |   pmei < 72.5
|   |   |   |   |   |   |   pcne < 48.5
```

Figura 7 Técnica de clasificación (*Classifying*) utilizando el algoritmo *RandomTree* de WEKA

Al contrario del primer modelo, el modelo generado con el algoritmo *RandomTree* (figura 7) es totalmente confiable al contar con un 100% de clasificación correcta; no obstante, en general no genera reglas interesantes ya que los grupos incluidos en el modelo son de tamaño pequeño, lo que los hace insuficientes para ser considerados parte de una regla.

ISSN 2410-3977

Con base en lo anterior, más que optar por otro algoritmo, se optó por utilizar otra técnica para minar los datos, a decir, la técnica de *Clustering* en conjunto con el algoritmo *SimpleKMeans* para generar 10 grupos. En la figura 8 se muestra el resultado de la aplicación de esta técnica, cuya interpretación se comenta en la sección de conclusiones.

Clustered Instances		Cluster 0	Cluster 1
0 288 (5%) 1 956 (16%) 2 371 (6%) 3 242 (4%) 4 239 (4%) 5 578 (10%) 6 515 (9%) 7 987 (17%) 8 331 (6%) 9 1342 (23%)	**sexo** **pcne** **nom_proc**	H 37.6354 *COLEGIO DE BACHILLERES DEL ESTADO DE ZACATECAS*	M 43.7385 *CENTRO DE BACHILLERATO TECNOLÓGICO AGROPECUARIO*

Cluster 2	Cluster 3	Cluster 4	Cluster 5
H 77.8976 *ESCUELA PREPARATORIA REGIONAL, U.DEG.*	H 66.9587 *COLEGIO DE ESTUDIOS CIENTÍFICOS Y TECNOLÓGICOS DEL ESTADO DE AGUASCALIENTES*	H 62.523 *COLEGIO DE BACHILLERES DEL ESTADO DE ZACATECAS*	H 57.6955 *ACUERDO 286*

Cluster 6	Cluster 7	Cluster 8	Cluster 9
H 64.5922 *CENTRO DE BACHILLERATO TECNOLÓGICO INDUSTRIAL Y DE SERVICIOS*	M 59.9382 *CENTRO DE BACHILLERATO TECNOLÓGICO INDUSTRIAL Y DE SERVICIOS*	H 51.1631 *ESCUELA PREPARATORIA REGIONAL, U.DEG.*	M 61.6744 **CLAVE NO RECUPERADA**

Figura 8 Técnica *Clustering* utilizando el algoritmo *SimpleKMeans* de WEKA

Conclusiones

En este artículo se presentaron los resultados preliminares de un estudio que se está llevando a cabo sobre la extracción de patrones y reglas de las bases de datos del EXANI-II en el Estado de Aguascalientes, utilizando para ello técnicas de minería de datos.

En principio fue necesario definir las variables de interés dentro de un conjunto de 98 variables que considera el EXANI-II y preparar los datos para poder minarlos, lo cual incluyó un proceso de limpieza y transformación de los mismos a un formato propio de la herramienta utilizada. Según los resultados obtenidos hasta el momento, existen tres instituciones que presentan áreas de oportunidad de refuerzo de conocimientos para sustentar el EXANI-II: el *Colegio de Bachilleres del Estado de Zacatecas*, el *Centro de Bachillerato Tecnológico Agropecuario*, en su población femenina particularmente, y la *Escuela Preparatoria Regional de la U.de G.*

Lo anterior podría sugerir que es necesario poner más atención a los estudiantes que proceden de otros Estados (Jalisco y Zacatecas), así como a aquellos que proceden de instituciones de corte agropecuario. Esta y otras situaciones deben ser analizadas con mayor detalle con base en otros modelos que afinen estas conclusiones y generen nuevas en pro de una efectiva definición de acciones para el mejoramiento de la calidad educativa.

Referencias

Bresfelean V. P. “Analysis and Predictions on Student’s Behaviour using Decision Trees in Weka Environment”, Babes- Bolyai University, Cluj-Napoca/Romania, 2007.

Cheewaprakobkit P. “Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Programs”, 2013.

CENEVAL Centro Nacional de Evaluación para la Educación Superior, A.C., Examen EXANI-II, http://www.ceneval.edu.mx/ceneval-web/content.do?page=1738, página revisada el 20 de abril de 2015.

ISSN 2410-3977

Kumar, S. A. and M.N. Vijayalakshmi. "Efficiency of Decision Trees in predicting Student's Academic Performance", 2011.

Pal A. K. & S. Pal. "Classification Model of Prediction for Placement of Students", 2013.

Ramanathan, L., S. Dhanda and S. Kumar. "Predicting students' performance using modified ID3 algorithm," *Inter. J. Eng. Tech.*, vol. 5, no. 3, pp. 2491-2497, June-July 2013.

The University of Waikato, Weka 3: Data Mining Software http://www.cs.waikato.ac.nz/ml/weka/, página revisada el 20 de abril de 2015.

ISSN 2410-3977