Artículo **Revista de Docencia e Investigación Educativa**
Marzo 2016 Vol.2 No.3 35-40

Extracción de reglas a partir de conjuntos de datos del EXANI-II mediante el clasificador J48 de WEKA

LUNA-RAMÍREZ, Enrique*†, HERNÁNDEZ-CHESSANI, David', SORIA-CRUZ, Jorge y CRUZ-VALENZUELA, Roberto.

Instituto Tecnológico El Llano Aguascalientes. Km. 18 Carr. Ags.-S.L.P. El Llano Aguascalientes. C.P. 20330
´Universidad Tecnológica de Ags. Blvd. Juan Pablo II 1302, Frac. Ex Hacienda la Cantera. Ags., Ags. C.P. 20206

Recibido Enero 7, 2016; Aceptado Marzo 9, 2016

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, particularmente mediante la aplicación de la técnica C4.5, conocida como clasificador J48 en la herramienta especializada WEKA, se encontraron reglas interesantes en conjuntos de datos del EXANI-II, que se espera sirvan de apoyo en 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. La validación de los modelos generados para la extracción de reglas se realizó mediante la prueba llamada *cross-validation*.

Minería de datos educativa, técnica C4.5

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, particularly by applying the C4.5 technique, known as J48 classifier in the specializied tool WEKA, interesting rules were found from EXANI-II data sets, which are expected to be useful in identifying those factors that impact negatively the academic performance of senior high students, as well as the definition of strategies to reinforce the weak identified aspects. The validation of the generated models for rule extraction was carried out through the so-called *cross-validation* test.

Educational data mining, C4.5 technique

Citación: LUNA-RAMÍREZ, Enrique, HERNÁNDEZ-CHESSANI, David', SORIA-CRUZ, Jorge y CRUZ-VALENZUELA, Roberto. Extracción de reglas a partir de conjuntos de datos del EXANI-II mediante el clasificador J48 de WEKA. Revista de Docencia e Investigación Educativa 2016, 2-3: 35-40

* 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, 2016) 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.

De manera particular, en la figura 1 se muestra la cantidad de alumnos que sustentaron el EXANI-II en el Estado de Aguascalientes, cuyos resultados fueron analizados mediante WEKA (The University of Waikato, 2016).

Selected attribute

Name: genero Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	Hombre	7197
2	Mujer	9555

Figura 1 Número de estudiantes que sustentaron el EXANI-II en el Estado de Aguascalientes

Es importante señalar que los datos de esta población requirieron ser preprocesados como un paso previo a su análisis, lo cual consistió básicamente en eliminar errores y redundancias presentes en los datos originales, así como recodificar algunas variables para poder manipularlas de manera más conveniente. Así, una vez preprocesados, los datos fueron ajustados al formato de WEKA para comenzar a generar modelos.

Revisión de literatura

Algunos de los trabajos más importantes realizados en torno al tópico de la *minería de datos educativa* que utilizan el clasificador J48 para descubrir conocimiento fueron descritos por Luna-Ramírez *et al.* (2015). En su artículo, los autores describen los trabajos de Bresfelean (2007), orientado a predecir la elección de carrera profesional de estudiantes de diferentes especialidades, de Cheewaprakobkit (2013), orientado a identificar "estudiantes débiles", de manera que el desempeño académico de tales estudiantes pueda ser mejorado, de Kumar y Vijayalakshmi (2011), orientado a predecir el desempeño de estudiantes en ciertos tipos de evaluación, de Pal y Pal (2013), orientado a predecir la colocación de estudiantes, y de Ramanathan *et al.* (2013) orientado a analizar el desempeño académico de estudiantes. No obstante, posterior al citado artículo, durante el año 2015, ocurrieron otras investigaciones relacionadas, descritas a continuación.

Pradeep *et al.* (2015) llevaron a cabo un estudio para predecir la reprobación y deserción de estudiantes "débiles", para lo cual analizaron 670 registros de estudiantes, con 57 atributos, entre los años 2011 y 2013, en una reconocida escuela en Kerala, India. En su estudio, los autores utilizaron diversas técnicas de clasificación de WEKA tales como las reglas de inducción y los árboles de decisión.

Pruthi y Bhatia (2015) utilizaron árboles de decisión y técnicas de agrupación, incluidas en WEKA, para predecir la colocación laboral de estudiantes egresados de carreras afines a las Tecnologías de Información y Comunicación, basándose en su historial académico. Así, los autores lograron predecir el tipo de compañía un egresado se colocaría, ya fuera consultora o desarrolladora, con una precisión del 95%. De hecho, el nombre de la compañía también fue predicho, con una precisión del 62%.

ISSN-2444-4952

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. 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, se determinó 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 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, que contempla la extracción, transformación y carga de los datos en un repositorio único, y su posterior explotación mediante herramientas especializadas, aunque a una escala menor, dado que en un Data Warehouse existen datos de diversos 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 algoritmos reconocidos como efectivos en diversas situaciones tales como el *J48* de WEKA para clasificar los datos mediante arboles de decisión.

ISSN-2444-4952

Los modelos generados fueron validados mediante la prueba llamada *cross-validation*. Para ello, se generaron modelos preliminares que dieron la pauta para analizar de manera posterior el desempeño académico por sectores (Luna-Ramírez *et al.*, 2015).

Como una etapa final, cuando se cuente con el conocimiento suficiente, este trabajo concluirá con la definición de estrategias y acciones para corregir deficiencias académicas que hayan sido detectadas, esto con el apoyo de expertos en educación.

Resultados

Con base en un análisis exploratorio de los datos preprocesados del EXANI-II y algunos modelos preliminares generados a partir de los mismos (Luna-Ramírez *et al.*, 2015), en una etapa posterior se lograron obtener algunas reglas mediante el clasificador J48 de WEKA y la prueba de validación denominada *cross-validation*. Estas reglas son presentadas en esta sección a continuación.

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic <= 65
|   regimen = Federal: Mujer (1610.0/68.0)
```

Figura 2 Regla relativa a mujeres que provienen de régimen federal

En la figura 2, se muestra una regla que indica que las mujeres que provienen del régimen federal suelen obtener una calificación máxima de 65 en la prueba de tecnologías de la información y comunicación (tic). De acuerdo al indicador de datos mal clasificados, esta regla tiene un error del 4.2%.

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic <= 65
|   regimen = Pública
|   |   nom_proc = CENTRO DE BACHILLERATO
|   |   |          TECNOLÓGICO INDUSTRIAL
|   |   |          Y DE SERVICIOS
|   |   |   prlm <= 40: Mujer (286.0/2.0)
```

Figura 3 Regla relativa a las mujeres que provienen de un CBTIS

En la figura 3, se muestra una regla que indica que las mujeres provenientes de CBTIS, que obtienen una calificación máxima de 65 en la prueba de tic, suelen obtener una calificación máxima de 40 en la prueba de razonamiento lógico matemático. De acuerdo al indicador de datos mal clasificados, esta regla tiene un error del 0.7%.

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic > 65
|   regimen = Pública
|   |   ptic > 80
|   |   |   nom_proc = CENTRO DE BACHILLERATO
|   |   |   |          TECNOLÓGICO INDUSTRIAL
|   |   |   |          Y DE SERVICIOS
|   |   |   |   pcne > 79: Hombre (127.0/7.0)
```

Figura 4 Regla relativa a los hombres que provienen de un CBTIS

En la figura 4, se muestra una regla que indica que los hombres provenientes de CBTIS, que obtienen una calificación mínima de 80 en la prueba de tic, suelen obtener una calificación mínima de 79 en el examen global de selección. De acuerdo al indicador de datos mal clasificados, esta regla tiene un error del 5.5%.

ISSN-2444-4952

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic <= 65
|   regimen = Pública
|   |   nom_proc = ESCUELA NORMAL DE
                   AGUASCALIENTES:
                   Mujer (128.0)
```

Figura 5 Regla relativa a mujeres que provienen de la Escuela Normal

En la figura 5, se muestra una regla que indica que las mujeres provenientes de la Escuela Normal suelen obtener una calificación máxima de 65 en la prueba de tic. De acuerdo al indicador de datos mal clasificados, esta regla no tiene error, es decir, es 100% confiable.

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic > 65
|   regimen = Pública
|   |   ptic <= 80
|   |   |   nom_proc = COLEGIO NACIONAL DE
|   |   |   |          EDUCACIÓN PROFESIONAL
|   |   |   |          TÉCNICA
|   |   |   |   prlm > 55
|   |   |   |   |   prv <= 70:
                      Hombre (103.0/1.0)
```

Figura 6 Regla relativa a los hombres que provienen del Colegio Nacional de Educación Profesional Técnica

En la figura 6, se muestra una regla que indica que los hombres provenientes del Colegio Nacional de Educación Profesional Técnica, con calificación mayor a 55 en la prueba de razonamiento lógico matemático y menor a 70 en la prueba de razonamiento verbal, suelen obtener una calificación entre 65 y 80 en la prueba de tic.

De acuerdo al indicador de datos mal clasificados, esta regla tiene un error del 1%.
En la figura 7, se muestra una regla que indica que las mujeres provenientes del Colegio de Bachilleres del Estado de Zacatecas, con calificación mayor a 30 en la prueba de español, pero menor a 35 en la prueba de razonamiento lógico matemático, suelen obtener una calificación menor a 44 en el examen global de selección. De acuerdo al indicador de datos mal clasificados, esta regla tiene un error del 12%.

Classifier output

```
Test mode:10-fold cross-validation

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

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

ptic <= 65
|   regimen = Pública
|   |   nom_proc = COLEGIO DE BACHILLERES DEL
|   |   |          ESTADO DE ZACATECAS
|   |   |   prlm <= 35
|   |   |   |   pesp > 30
|   |   |   |   |   pcne <= 44: Mujer (65.0/8.0)
```

Figura 7 Regla relativa a mujeres que provienen del Colegio de Bachilleres del Estado de Zacatecas

Conclusiones

En este artículo se presentaron diversas reglas encontradas en conjuntos de datos del EXANI-II, extraídas mediante el clasificador J48 de WEKA. Los modelos de los cuales se extrajeron las reglas fueron validados mediante la prueba denominada *cross-valiation*, ténica que permite garantizar que los resultados obtenidos son independientes de la partición entre los datos de entrenamiento y los datos de prueba.

ISSN-2444-4952

Para poder llevar a cabo esta investigación, fue necesario definir primero las variables de interés dentro del conjunto de 98 variables que considera el EXANI-II y, posteriormente, preprocesar los datos para poder minarlos, lo cual incluyó un proceso de limpieza de datos y su transformación al formato de WEKA.

En el estudio realizado, se detectaron reglas sobre estudiantes de ambos sexos, ninguna de las cuales incluyó al régimen privado, lo que invita a pensar que el régimen público (incluido el federal) continúa siendo de mayor trascendencia.

Se detectaron reglas que muestran de manera clara algunas correlaciones entre las diferentes áreas que son examinadas en el EXANI-II, lo cual puede servir de apoyo para definir nuevas estrategias para mejorar la calidad educativa a nivel medio superior.

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 2016.

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

Luna-Ramírez, E., Correa-Villalón, C., Velarde-Martínez, A., Hernández-Chessani, D. "Análisis de los resultados del EXANI-II en el Estado de Aguascalientes mediante técnicas de minería de datos", Revista de Sistemas y Gestión Educativa, Vol. 2, Num. 2, pp. 206-213, Enero-Marzo 2015.

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

Pradeep, A.; S. Das and J. J Kizhekkethottam. "Students dropout factor prediction using EDM techniques", International Conference on Soft-Computing and Networks Security, pp. 1-7 2015.

Pruthi, K. and P. Bhatia. "Application of Data Mining in predicting placement of students", International Conference on Green Computing and Internet of Things, pp. 528 – 533, 2015.

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 2016.

ISSN-2444-4952