Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE

Dapozo, Gladys; Porcel, Eduardo; López, María V.; Bogado,Verónica; Bargiela, Roberto

Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste. 9 de Julio nº 1449. CP: 3400. Corrientes. Argentina.
TE: (03783) 423126 gndapozo@exa.unne.edu.ar; eporcel@exa.unne.edu.ar; mvlopez@exa.unne.edu.ar

RESUMEN

El sistema preuniversitario argentino tiene serias deficiencias, y una de las consecuencias se manifiesta en que el piso cognitivo y actitudinal con el que ingresan los alumnos a la Universidad es muy bajo y atenta contra el rendimiento académico de los mismos, además de contribuir a la extensión de la duración real de las carreras. La Minería de Datos abarca una variedad de métodos estadísticos y computacionales para investigar la existencia de relaciones y patrones de comportamiento en almacenamientos electrónicos de datos. El Software Libre resulta más adecuado que el software propietario para entornos académicos al ser más fiable, robusto y seguro y de reducido costo. En este trabajo se presenta un estudio a través de técnicas de minería de datos que permiten determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste (FACENA-UNNE). Se llevó a cabo un estudio comparativo de diferentes algoritmos clasificadores disponibles en el software Weka, de libre distribución, y se seleccionó el que ofrecía mejores resultados.

Palabras clave: Minería de datos. Herramienta de software libre. Rendimiento académico de alumnos universitarios.

1. INTRODUCCIÓN

El sistema preuniversitario argentino tiene serias deficiencias, y una de las consecuencias se manifiesta en que el piso cognitivo y actitudinal con el que ingresan los alumnos a la Universidad es muy bajo y atenta contra el rendimiento académico de los mismos, además de contribuir a la extensión de la duración real de las carreras. Todo ello constituye un indicador desfavorable para la fórmula polinómica de asignación de recursos económicos a la Universidad.

Con respecto a los estudios de rendimiento académico, se sabe que se trata de un problema multifacético resultante de numerosas causas y condicionantes económicos, culturales, políticos, demográficos. Esta complejidad exige que el problema sea abordado considerando la totalidad de la información de los alumnos que las instituciones universitarias disponen en formato electrónico [1].

La Minería de Datos (Datamining), o Descubrimiento de Conocimiento en Bases de Datos, abarca una variedad de métodos estadísticos y computacionales para investigar la existencia de relaciones y patrones de comportamiento en almacenamientos electrónicos de datos. Relaciones y patrones emergentes pueden sugerir al investigador explicaciones causales que puedan ser verificadas posteriormente o bien pueden sugerir estrategias de acción para lograr ciertos objetivos de cambio [7].

Esta tecnología emergente combina los análisis estadísticos, máquina de aprendizaje y la gestión de las bases de datos para extraer información de voluminosas tablas de datos [9].

La implementación de las técnicas de minería no implica siempre grandes inversiones. Generalmente se recurren a programas costosos para este tipo de tareas. Sin embargo, existen herramientas más simples y menos costosas que pueden brindar las mismas prestaciones para el conjunto de datos con que se cuenta [4].

Por tales motivos, se debe realizar un análisis de los sistemas implementados y de los datos, determinar la técnica de Datamining que más se adecue y, luego, elegir la herramienta, si es que existiere una, o programarla en caso contrario. Esto permite dotar a la organización de un potente Datamining y tener

un costo menor que adquirir una solución propietaria de grandes dimensiones y compleja, donde el tiempo invertido en aprendizaje puede ser demasiado [6].

El papel del software libre (SL) en la universidad no se reduce a la disponibilidad de una sofisticada plataforma de desarrollo tecnológico. Por el contrario, es un fenómeno de gran calado cuyas dimensiones éticas y sociales pueden transformar el marco académico, haciéndolo más democrático, participativo y viable en términos financieros [2].

Propiciada por las Tecnologías de la Información y de la Comunicación (TICs), surge una nueva ecología del conocimiento que consiste en otras formas epistémicas y metodologías de conocimiento que definen el tránsito de una sociedad de la información a una sociedad del conocimiento, donde ese saber que fluye por las venas del tejido social se verticaliza, se transforma cualitativamente en su recurso fundamental de supervivencia. Y es aquí donde el SL tiene un papel fundamental, pues su metodología se corresponde con una revolución organizacional fundamental: el paso de los modelos jerárquicos a los modelos en red, a las organizaciones e instituciones flexibles y dinámicas que se adaptan con mayor facilidad a su medio ambiente.

A nivel institucional, tres ventajas son claves: En primer lugar, el SL es más adecuado que el software propietario para entornos académicos al ser más fiable, robusto y seguro. En segundo lugar, su reducido costo permite localizar recursos financieros en otras áreas de las universidades (infraestructuras, becas, apoyo a la investigación, etc.). En tercer lugar, al demandar menores recursos computacionales, se extiende la vida útil de los equipamientos informáticos, evitando ciclos rápidos de obsolescencia y optimizando así las inversiones.

A nivel académico, el SL refleja mucho mejor los valores tradicionales de la investigación universitaria desde su propia definición de "libre": libertad para analizar cómo trabaja un programa y adaptarlo a nuestras necesidades, libertad para mejorar un programa y compartir con otros las adaptaciones, beneficiando así a toda la comunidad.

A nivel metodológico, se quiebra el paradigma neoliberal de maximización del beneficio individual, sustituyendo la competición por la sinergia, esto es, por la convergencia de esfuerzos individuales en pro de un objetivo común.

En el movimiento de SL, la interactividad y la participación activa se revelan como las reglas básicas del juego. Con su llegada, los métodos de desarrollo de software y de acceso y distribución de la información cambiaron radicalmente. Para el mundo de la formación, especialmente, el de la universidad, esta característica posee una tremenda carga transgresora con respecto a los modos clásicos de aprender, producir y distribuir en este ámbito [2].

La clasificación de las técnicas de la minería de datos se divide en dos categorías: supervisadas y no supervisadas [3] [10]. Las primeras predicen los valores de un atributo etiqueta u objetivo con la ayuda de los valores de otros atributos, por lo que van a estar dirigidas a la clasificación y a los sistemas de predicción. En el caso de las técnicas no supervisadas, a partir de un conjunto de datos disponible se persigue encontrar relaciones entre los atributos, patrones habituales de comportamiento, desconocidos antes del análisis, de ahí que a este tipo de técnicas también se les llame de descubrimiento del conocimiento [8].

En este estudio, el trabajo está encaminado a comparar algoritmos supervisados a través de clasificadores.

El objetivo de este trabajo es presentar un estudio a través de técnicas de minería de datos que permitan determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste (FACENA-UNNE). Se llevó a cabo un estudio comparativo de diferentes algoritmos clasificadores disponibles en el software Weka, de libre distribución, y se analizaron los resultados que se obtuvieron como resultado de la aplicación de cada uno de ellos.

2. MATERIALES Y TÉCNICAS A EMPLEAR

En este trabajo se utilizó la herramienta Weka (Waikato Environment for Knowledge Analysis) de la Universidad de Waikato, software que se encuentra de manera gratuita en el sitio oficial de esta institución en Internet

y contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas [5].

Los datos utilizados en este análisis fueron obtenidos de un almacén de datos que integra toda la información sistematizada de los alumnos de la Facultad de Ciencias Exactas de la UNNE. El mismo contiene los datos particulares y socio económicos que se registran en el ingreso, los datos de todas las actividades académicas, como asignaturas cursadas y rendidas, trámites de reinscripción y readmisión, reconocimiento de materias y datos del egreso o trámite de graduación [1].

El almacén de datos está contenido en una base de datos Access. A través de consultas SQL se obtuvo el conjunto de datos para este análisis particular, integrando los datos de la tabla Ingresantes que posee la información socioeconómica y del nivel educativo previo del alumno y los datos de la tabla Situación Académica, que contiene el registro de todas las actividades de los alumnos.

Luego se seleccionaron los alumnos que pertenecen a la carrera Licenciatura en Sistemas de Información que rindieron exámenes finales de las materias que corresponden al primer año en fechas correspondientes al año del ingreso. Con esta información, para cada alumno se calculó: la cantidad de exámenes finales rendidos (número de intentos), la cantidad de exámenes finales aprobados y la cantidad de exámenes finales desaprobados. En función de estos valores, se generaron las tres categorías que identificarán a los alumnos que: 1) en el año de ingreso no rindieron ninguna materia, 2) rindieron pero no aprobaron ninguna y 3) rindieron y aprobaron por lo menos una materia. La consulta resultante se exportó a una planilla de cálculo.

El archivo fue formateado para cumplir con las restricciones del programa Weka que fue utilizado para el procesamiento de los datos, y contiene 2887 registros con las siguientes variables referidas a los alumnos: año de ingreso (ANIO), sexo (SEXO), estado civil (CIVIL), situación laboral del alumno (SILAAL), grado de instrucción del padre (GRAINSPA), situación laboral del padre (SILAPA), categoría ocupacional del padre (CAOCPA), grado de instrucción de la madre (GRAINSMA), situación laboral de la madre (SILAMA), categoría ocupacional de la madre (CAOCMA), título secundario (TITULO), dependencia del establecimiento secundario (DEPENSEC) y categoría de alumno según la cantidad de materias aprobadas en primer año (CAT_ALUMNO).

Para facilitar la comprensión de las salidas y gráficos de los algoritmos de Weka, conviene que las variables sean de tipo cualitativo. Por tanto, se codificaron todas las variables como cualitativas o nominales, lo cual requirió un trabajo previo realizado con planillas de cálculo. Luego se procedió a la construcción del archivo en formato AARF, empleando un editor de textos.

La variable o atributo conocido a predecir en este trabajo está representada por CAT_ALUMNO. La misma comprende tres categorías de alumnos, según su rendimiento académico durante el primer año, relacionado con los intentos y resultados de exámenes finales: 1 (no se presentó a rendir nunca), 2 (se presentó a rendir pero no aprobó ninguna materia) y 3 (aprobó una o más materias).

2. RESULTADOS Y DISCUSIÓN

En la Figura 1 se muestra a través del Explorer de Weka la composición del conjunto de datos y el número de registros por categoría de la variable CAT_ALUMNO. Por su parte, en la Figura 2 se visualiza el número de registros por año de ingreso (ANIO), y la proporción de alumnos de categorías 1, 2 y 3 en cada año. Se observa que en los años 2004 y 2005 ha aumentado la proporción de alumnos que no rinden ninguna materia en el primer año y ha disminuido la proporción de alumnos que aprueban al menos una materia durante el primer año.

Finalmente, en la Figura 3 se ilustra el número de registros para las distintas variables en función de las categorías de CAT_ALUMNO.

A continuación, se probaron diferentes algoritmos clasificadores del software Weka, para seleccionar aquél que con un menor error construyese un clasificador para la predicción de la categoría de alumno según su comportamiento durante el primer año (CAT_ALUMNO).

Los mejores resultados fueron obtenidos con el clasificador **Logistic** (Figura 4), el cual permite estimar y luego emplear modelos de regresión logística múltiple. En el estudio de estos datos se obtuvieron resultados con mediano grado de precisión, ya que el error del clasificador fue de 36,024%, y el porcentaje de

instancias clasificadas correctamente fue de 63,97%.

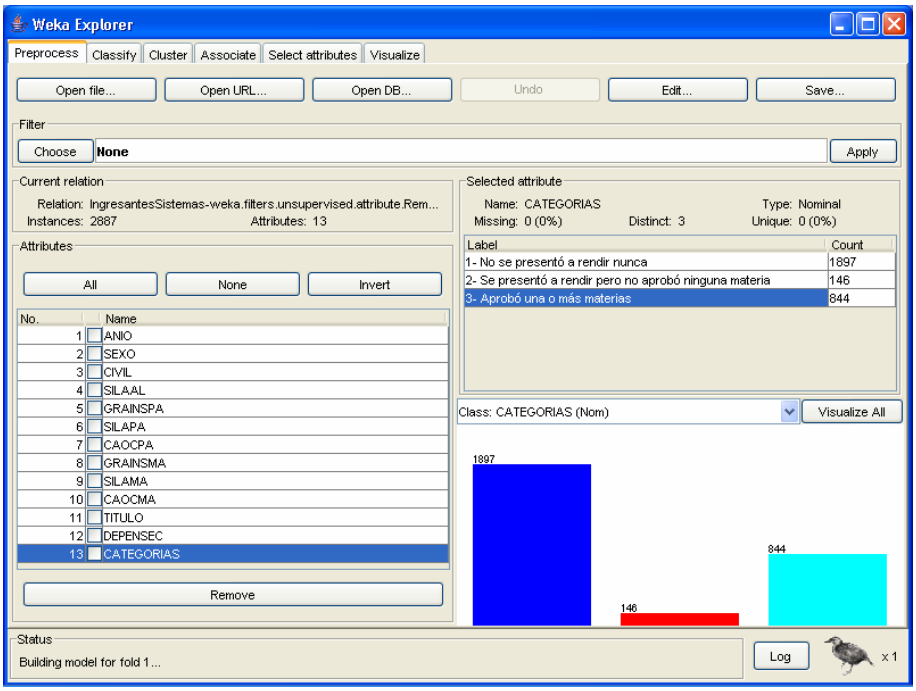

Figura 1. Composición de la primera base de datos estudiada a través de Weka y visualización del número de registros en función de las categorías de CAT_ALUMNO

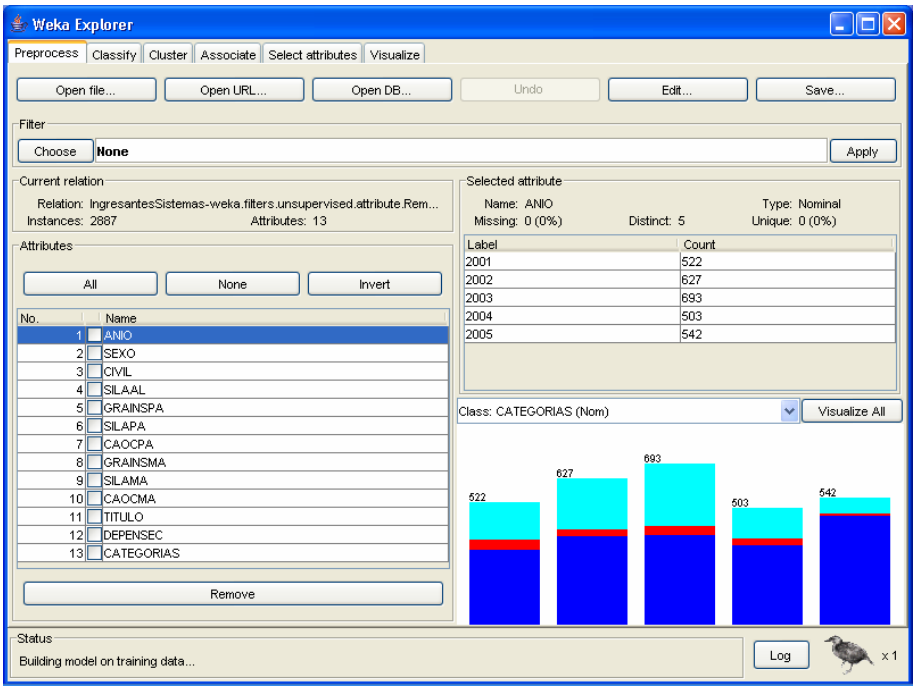

Figura 2. Visualización del número de registros de la variable Año de ingreso (ANIO) en función de la variable CAT_ALUMNO

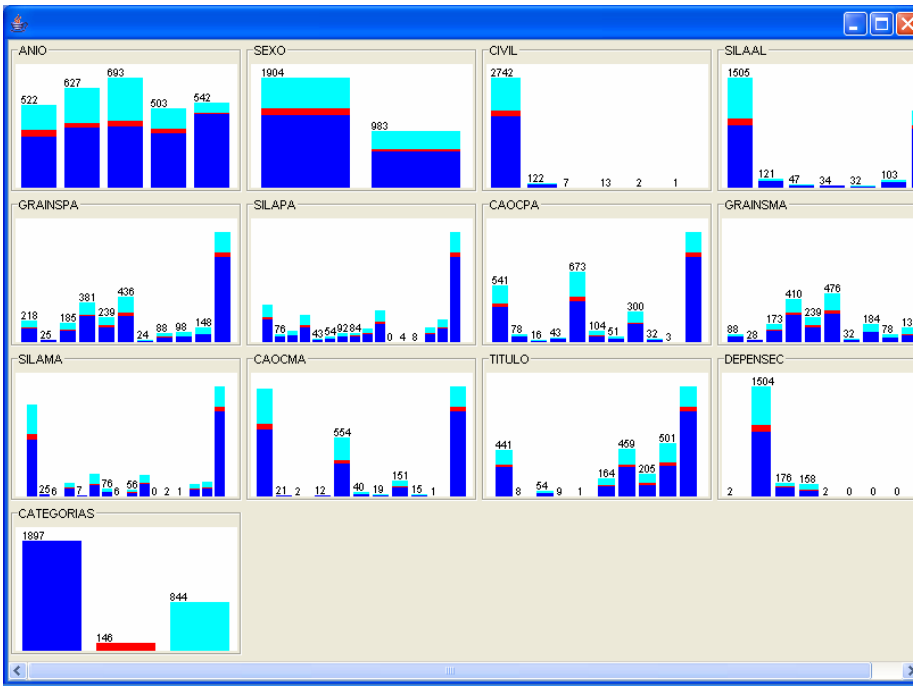

Figura 3. Visualización del número de registros de cada variable en función de la variable CAT_ALUMNO

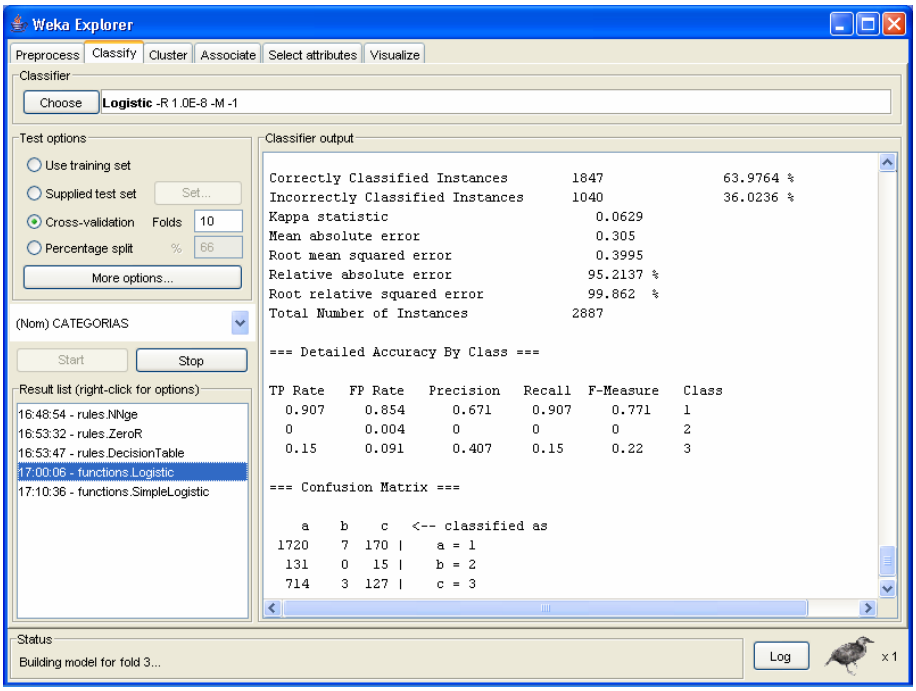

Figura 4. Parte de la salida obtenida mediante el clasificador Logistic de Weka

4. CONCLUSIONES

A través del uso de la minería de datos se han probado diferentes algoritmos clasificadores disponibles en el software Weka de libre distribución, con el objeto de encontrar un clasificador que predijera los valores de la variable CAT_ALUMNO, que describe la categoría de alumno según los intentos realizados y resultados obtenidos en los exámenes finales de las materias del primer año de la carrera. Esto permitió estimar el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de FACENA-UNNE.

Si bien no se encontrado un clasificador que prediga la variable en estudio con un alto grado de precisión, el uso de la herramienta informática Weka permitió realizar un análisis

descriptivo de los datos mediante gráficos, de modo sencillo. Sin embargo, se ha observado que, a pesar de que este software ofrece muchos algoritmos para la construcción de clasificadores, carece de una documentación o ayuda adecuada.

Los autores proponen continuar con el estudio y prueba de los algoritmos que ofrece el software Weka (más de 20), que podrían mejorar los resultados obtenidos en este trabajo.

5. REFERENCIAS

[1] Dapozo, G., Porcel, E. "Metodología de integración de datos para apoyar el seguimiento y análisis del rendimiento académico de los alumnos de la FACENA". Comunicaciones Científicas y Tecnológicas de la UNNE 2005. http://www.unne.edu.ar/Web/cyt/com2005/8-Exactas/E-032.pdf.

[2] Bustamante Donas, Javier. "El software libre y la universidad". http://www.libroblanco.com/html/modules.php?op=modload&name=News&file=article&sid=164&mode=thread&order=0&thold=0.

[3] Herschkowitz, D. and J. P. Nadal. "Unsupervised and supervised learning: Mutual information between parameters and observations". Physical Review E, The American Physical Society, Volume 59, Number 3, March, pp. 3344-3360, 1999. http://www.menem.com/~ilya/digital_library/learning/hershkowitz-nadal.pdf.

[4] Kleissner, C. "Datamining for the enterprise".System Sciences, Proceedings of the Thirty-First Hawaii International Conference on,Volume 7, 6-9 Jan., pp. 295-304, 1998.

[5] Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/

[6] Redondo, Juan U. "Cómo sacar partido con inteligencia de los datos". http://c.microsoft.com/trans_pixel.asp

[7] Sananes, Marta; Torres, Elizabeth; Sinha, Surendra P. y Nava Puente, Luis. "Búsqueda y caracterización de subgrupos de pobreza mediante la aplicación de algunas técnicas de minería de datos". Instituto de Estadística Aplicada y Computación, Universidad de Los Andes, Mérida, Venezuela. Escuela de Estadística, Universidad de Los Andes, Mérida, Venezuela.

[8] Segrera, Saddys; Moreno, María N.; Miguel, Luis A. "Aplicación de la minería de datos en la evaluación de la aptitud física de las tierras para el cultivo de la caña de azúcar". Dept. de Informática. Instituto Nacional de Investigaciones de la Caña de Azúcar. Ciudad de la Habana. Dept. de Informática y Automática. Facultad de Ciencias. Univ. de Salamanca. Salamanca.

[9] Thuraisingham, B. "A primer for understanding and applying Datamining. IT Professional". Volume 2, Issue 1, Jan.-Feb., pp. 28-31, 2000.

[10] Weiss, S.M. and N. Indurkhyya. "Predictive Datamining. A Practical Guide". Morgan Kaufmann Publishers, San Francisco, 1998.