Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. Un caso de estudio

Laura Lanzarini[1], Emilia Charnelli[2], Guillermo Baldino[3], Javier Díaz [2]

[1] Instituto de Investigación en Informática LIDI (III-LIDI).Facultad de Informática. UNLP.
[2] Laboratorio de Investigación en Nuevas Tecnologías Informáticas (LINTI). Facultad de Informática. UNLP.
[3] Laboratorio de Innovaciones en Sistemas de Información (LINSI). Dpto de Sistemas. UTN.

laural@lidi.info.unlp.edu.ar, mcharnelli@linti.unlp.edu.ar, gbaldino@linsi.edu.ar, javierd@info.unlp.edu.ar

Resumen

La Minería de Datos Educativa reúne a los distintos métodos que permiten extraer información novedosa y útil a partir de grandes volúmenes de datos provenientes de contextos educativos.

El presente trabajo describe el proceso de identificación, a través de técnicas de visualización, de las características más relevantes en lo que se refiere al rendimiento académico de los alumnos de la Facultad de Informática de la Universidad Nacional de La Plata. Este es un paso inicial que ejerce una gran influencia en la eficiencia y eficacia de los métodos que permiten modelar la información ya que los resultados a obtener mejoran al reducir la dimensión del problema. Esto último redunda en una representación más clara y simple de la información disponible. Para lograrlo, en este artículo se propone analizar y aplicar, luego del preprocesamiento de los datos, diferentes visualizaciones de los atributos sobre las clases o respuestas esperadas. Con este enfoque se espera generar una metodología de trabajo que ofrezca resultados fáciles de usar e interpretar. Su aplicación a la información correspondiente a alumnos regulares y no regulares de la UNLP permitió establecer relaciones interesantes acerca del desempeño académico de los alumnos. Esto último impacta directamente en las condiciones por las cuales abandonan sus estudios universitarios.

Palabras claves: mineria de datos educativa, visualizacion, educacion, atributo de selección, rendimiento académico.

Abstract

Educational Data Mining collects the various methods that allow extracting novelty and useful information from large data volumes in educational contexts.

This paper describes the process used to, through advanced visualization techniques, identify the most relevant characteristics in relation to student academic performance at the School of Computer Science of the National University of La Plata. This is the initial step that greatly affects the efficiency and efficacy of the methods that are used to model the information, since the results obtained improve when the dimension of the problem decreases. This in turn results in a clearer and simpler representation of the available information. To achieve this, we propose analyzing and applying, after a data pre-processing stage, various visualization methods for the attributes for the classes or expected responses. With this approach, we expect to develop a work methodology that offers results that can be easily used and interpreted. Its application to the information relating to regular and non-regular students at the UNLP allowed establishing interesting relationships in relation to student academic performance. This directly affects the reasons why students drop out from university.

Key words: Educational Data Mining, Visualization, Education, Attribute Selection, Academic Performance.

Introducción

En la actualidad, la mayoría de los procesos, ya sean industriales, académicos, de negocios o de servicios, cuentan con información histórica almacenada. El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes y difíciles de comprender y analizar. Distintas áreas han tratado de dar soluciones a este problema. Las técnicas de visualización a través de representaciones gráficas, algunas de las cuales son sumamente sofisticadas, han contribuido significativamente a la exploración y entendimiento de estos conjuntos de datos [1][2][3]. Por su parte, la Minería de Datos reúne un conjunto de técnicas capaces de modelizar y resumir la información, facilitando su comprensión y ayudando a la toma de decisiones en situaciones futuras [4][5][6].

El área educativa no escapa a esta realidad. Por lo general, los establecimientos disponen de información sumamente detallada de cada alumno pero carecen de modelos que les permitan describir de manera objetiva a sus estudiantes. Caracterizar a los estudiantes de una institución académica aporta información no trivial y de utilidad para la toma de decisiones, como por ejemplo, establecer políticas tendientes a mejorar el desempeño académico de los alumnos lo cual redundará en la reducción de la deserción universitaria.

El objeto de estudio presentado en este artículo es la Facultad de Informática de la UNLP. Dicha institución fue creada en 1999, aunque sus carreras de grado comenzaron en el año 1966 dentro de la Facultad de Ciencias Exactas. Actualmente, en dicha Facultad se dictan 3 carreras de grado, Licenciatura en Sistemas, Licenciatura en Informática y Analista Programador Universitario, y en conjunto con la Facultad de Ingeniería, dictan la carrera Ingeniería en Computación, con un promedio anual de aproximadamente 800 ingresantes.

Actualmente la problemática de deserción en las carreras de Informática forma parte de una situación a la cual se enfrentan tanto autoridades como docentes. Existen distintas herramientas, tales como becas, programas de tutorías y seguimientos por parte de los gabinetes pedagógicos para trabajar con alumnos que abandonan la carrera. Si bien se considera que estas herramientas son útiles e importantes, actúan en instancias en las cuales el alumno ya tomó la decisión de abandonar sus estudios.

La comunidad educativa coincide en la necesidad de hacer esfuerzos para revertir esta situación y cualquier tipo de medidas que se adopten deben estar basadas en información útil para la rápida toma de decisiones. Distintos autores han propuesto diferentes enfoques relacionados con la captación de estudiantes como en el análisis y detección de abandonos y también con la estimación de la duración de la carrera [7] [8] [9] [10] [11] [12] [13]. También hay autores que han estudiado cómo evoluciona el progreso de los alumnos durante sus estudios [14] [15].

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (Knowledge Discovery in Databases) y técnicas de Visualización aplicadas al análisis de la información disponible.

El proceso de KDD tiene como objetivo la detección automática de patrones sin necesidad de contar con una hipótesis especificada a priori. Sin embargo, su aplicación requiere identificar, en base al problema a resolver, cuál es la información sobre la que se va a trabajar y cuál es el tipo de modelo que se desea obtener.

En referencia a la información sobre la que se va a trabajar, este artículo propone una metodología de trabajo que utiliza visualizaciones de la información disponible para identificar los atributos que mejor caracterizan el avance académico de un alumno logrando reducir así la información a considerar. Esto permite enfocar el análisis en las características adecuadas y arribar a un perfil de alumno de fácil interpretación. Es sabido que el objetivo de una visualización es lograr una representación que ayude al usuario a interpretar un conjunto de datos y comunicar su significado [17]. Sin embargo, es común que la información que se desea representar no tenga una manifestación visual obvia. Ante esta situación el proceso de mapeo del conjunto de datos originales a la vista minable o información a procesar a través del método seleccionado, puede llegar a ser no trivial [16].

La UNLP utiliza el sistema SIU-Guaraní para la gestión académica de sus alumnos. Este sistema almacena los datos en una base de datos relacional. La información recolectada involucra a 5268 alumnos de la Facultad de Informática comprendido entre los años 2002 y 2012. Debido a la gran cantidad de datos que componen el dominio a trabajar se dificulta la identificación de patrones o relaciones existentes entre las opiniones de distintos sujetos. Por esto último, resulta de interés recurrir a técnicas objetivas que permitan identificar las características más relevantes. Sin embargo, antes de aplicar técnicas específicas de Minería de Datos y Visualización, es preciso verificar y preparar la información a fin de evitar inconsistencias.

Este trabajo está organizado de la siguiente forma: la sección 2 describe el preprocesamiento efectuado sobre los datos originales, la sección 3 muestra la selección de atributos relevantes a través de la generación de diferentes visualizaciones, la sección 4 muestra la construcción de un modelo a partir de los atributos seleccionados y los resultados obtenidos, mientras que en la sección 5 se presentan las conclusiones de este trabajo.

Preparación de los datos

Las primeras etapas del proceso de KDD involucran la comprensión del dominio y la recopilación de los datos. Los datos de los alumnos de la Facultad de Informática fueron recolectados del sistema SIU-Guaraní. Generalmente en la mayoría de las unidades académicas de la UNLP, la información personal de los alumnos es cargada por los mismos a la hora de inscribirse en una carrera universitaria. El cuestionario del sistema contiene información tanto sobre datos personales como laborales y se organiza como se observa en la Tabla 1.

1. Datos Personales (estado civil, familiares a cargo, con quien vive, etc.).
2. Financiamiento de estudios (familia, beca, trabajo).
3. Situación laboral (si busca trabajo, cuántas horas trabaja, relación con la carrera.
4. Situación padres (si viven, nivel de estudios y su actividad profesional).
5. Otros estudios.
6. Tecnología (si dispone de PC, acceso a Internet, etc.)
7. Nivel de idiomas.

Tabla 1. Estructura de la encuesta en SIU-Guaraní

Completadas las primeras etapas del KDD, se continúa con la etapa de preparación de los datos. Es necesario seleccionar y preparar el subconjunto de datos a utilizar. Esta fase cubre todas las actividades para construir el conjunto final de los datos que serán utilizados por las técnicas de modelado.

Atributos con datos inconsistentes

En la figura 1 se muestra una visualización que representa un diagrama de dispersión con las respuestas de los alumnos. Cada fila de la matriz representa a un alumno y cada columna una pregunta del cuestionario. El color oscuro significa que un alumno no respondió la pregunta. Se observa el patrón que siguen las preguntas no contestadas del SIU-Guaraní en los últimos años. La mayoría de las respuestas nulas se corresponden con las últimas preguntas del cuestionario web, algunas de las cuáles se tratan de otros estudios realizados, nivel de idiomas, uso de la PC e Internet, etc. En las primeras preguntas, se destacan las respuestas nulas de cantidad de familiares a cargo y relación del trabajo con la carrera, ya que más del 50 por ciento respondió que no trabaja y que vive con la familia.

Los atributos con más de un 80 por ciento de nulos fueron descartados para el análisis.

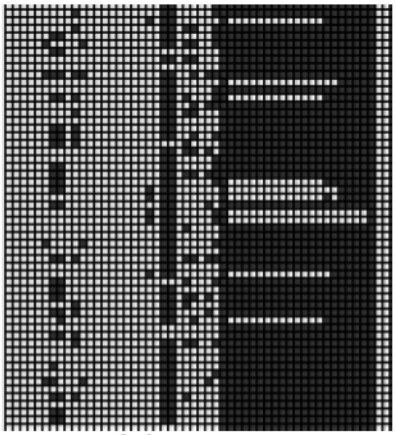

Figura 1. Patrón de las preguntas no contestadas

Atributos con datos no generalizables

Se eliminaron atributos no generalizables como el nombre del estudiante, el número de documento, el número de legajo y el número de inscripción.

Se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas incrementando así su capacidad predictiva. Por ejemplo, en los datos originales se registra colegio secundario con 453 valores diferentes, que representan los nombres de los colegios en los que los alumnos cursaron el nivel medio. Por lo que el nombre de la escuela fue reemplazado por dos atributos, uno que indica si la institución es pública o privada, y otro que indica si la institución es técnica o no.

Algo similar se hizo con el lugar de procedencia del alumno, reemplazándolo con el atributo que indica si es del interior o no, para este caso se consideró si las localidades estaban a más de 60 km de la ciudad de La Plata, lugar donde se encuentra dicha unidad académica.

Se redujo la cardinalidad de la actividad laboral de los padres con los valores más representativos en cada caso. En la Facultad de Informática-UNLP ser empleado representa la mayoría de los valores que toma el atributo actividad laboral tanto de la madre como del padre. Se redujo la cardinalidad a si tiene relación de dependencia o no, ya que tener personas a cargo, ser ama de casa o bien ser independiente se pueden unificar como actividades que no son dependientes.

Transformaciones realizadas

La creación de características consiste en generar nuevos atributos con el objetivo de mejorar la calidad y comprensión del conocimiento extraído. En esta dirección, se transformó la fecha de nacimiento por la edad de los alumnos. Por otro lado, a partir del año de egreso del secundario y del año de ingreso a la facultad, se generó un nuevo atributo que calcula esta diferencia.

Otras transformaciones que se realizaron tienen que ver con si trabaja y busca trabajo, cómo costea sus estudios, con quién vive. Se realizó la numerización de algunos atributos y la posterior normalización de su rango, de acuerdo a los requerimientos de las técnicas de Minería de Datos a utilizar. Se numerizó el máximo nivel de estudios de los padres y la cantidad de horas semanales que trabaja el alumno.

Además de los datos censales, se dispone de toda la información académica de los estudiantes.

Con el objetivo de analizar el avance de los alumnos en sus estudios y por cuestiones de simplicidad sólo se trabajó con la cantidad de finales aprobados por alumno al finalizar cada año durante los primeros 5 años de su vida universitaria. Se consideró que esta cantidad de años resulta representativa y coincide con la duración de las carreras de la Facultad. Los valores de estos atributos se obtienen al calcular para cada alumno la proporción de finales aprobados desde el inicio de su carrera hasta el final de cada año lectivo en relación a la cantidad total de materias según cada carrera como se observa en la siguiente ecuación

$$avance_i = \frac{f_i}{F} \qquad i=1..5 \qquad (1)$$

donde

- f_i es la cantidad total de materias que el alumno registra como aprobadas al finalizar el i-ésimo año.
- F es la cantidad total de materias de la carrera.
- $avance_i$ es un valor entre 0 y 1 que representa el avance que el alumno tiene en su carrera al finalizar el i-ésimo año.

Por último, se creó un campo que resume el estado académico del alumno, cuyo valor indica si se trata de un alumno regular o no, teniendo en cuenta las condiciones de pérdida de regularidad establecidas en la Facultad de Informática:

"Establecer la condición de regularidad para33 todos los alumnos de la Facultad (incluyendo ingresantes) con la Aprobación de 1 Examen Final o de 1 Cursada de Trabajos Prácticos (1 actividad positiva) durante el transcurso de los últimos 3 ciclos lectivos (1 de Marzo a 28 de febrero)".

Selección de atributos por medio de la Visualización

Las técnicas de Minería de Datos aplicadas sobre ejemplos de dimensión alta dan como resultado modelos complejos. Dependiendo de la técnica utilizada, datos con esta característica producen o bien árboles enormes o conjuntos de reglas con alta cardinalidad y antecedentes formados por un número importante de conjunciones [18] o funciones discriminantes difíciles de interpretar.

Para resolver este problema es preciso analizar, en forma previa a la construcción del modelo, cuáles son los atributos más representativos de la información disponible. Una vez seleccionados los atributos más relevantes, la técnica a utilizar verá simplificada su tarea y ofrecerá como resultado un modelo más sencillo y fácil de interpretar [19].

En el caso particular del problema a resolver en este artículo, la selección de características juega un rol fundamental ya que se espera poder identificar los atributos adecuados que permitan construir un modelo del avance académico de los alumnos por tratarse de una métrica estrechamente relacionada con la condición de regularidad.

Por lo tanto, luego de la etapa de preparación de los datos, se trabajó con técnicas de visualización para identificar dichos atributos.

Tras evaluar diferentes metodologías, se consideró utilizar la técnica de *coordenadas paralelas* ya que resulta adecuada para visualizar conjuntos de datos multidimensionales [20]. Informalmente, esta técnica de coordenadas paralelas consiste en asignarle a cada dimensión un eje y disponer estos ejes paralelamente en el plano. Además de ser una técnica apta para datos multidimensionales, es también apropiada para grandes conjuntos de datos [21].

Con el objetivo de analizar el avance de los alumnos en cada uno de los primeros 5 años de su vida universitaria se utilizaron los atributos creados según la ecuación (1) y se construyeron las gráficas que se observan en la figura 2 separando los alumnos regulares (figura 2.a) de los no regulares (figura 2.b). En cada caso, se buscó identificar 3 grupos de alumnos: los de mejor desempeño (línea con cruces), los de desempeño medio (línea con cuadrados) y los de bajo desempeño (línea sola).

En la figura 2, se optó por una visualización simplificada de la técnica coordenadas paralelas donde para cada atributo de cada grupo sólo se representa su valor promedio (línea central) y su desviación (zona sombreada que rodea a la línea central). De esta forma se tiene una representación más conceptual de cada grupo.

Luego, observando la figura 2 puede advertirse la relación que existe entre el rendimiento de los alumnos en los primeros cinco años y su condición de regularidad. Así también, se puede notar que se produce un punto de inflexión en el segundo año y a partir de ese momento una gran cantidad de alumnos detienen su progreso en la carrera.

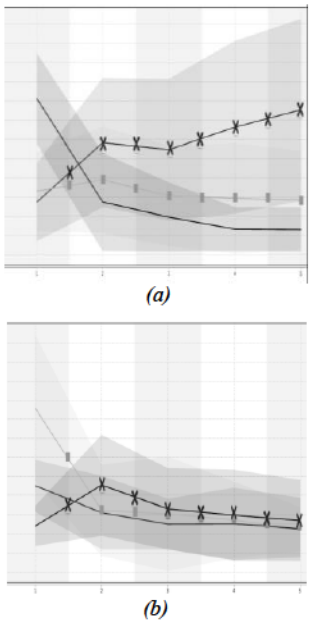

Figura 2 Gráfico de coordenadas paralelas simplificado donde sólo se representa la media y la desviación de los atributos avance$_i$ con i=1:5 para (a) Alumnos regulares (b) Alumnos no regulares

A partir de que el progreso de los alumnos está muy ligado a su condición de regularidad, se analizó la relación de cada uno de los 40 atributos según el progreso del alumno. El resultado obtenido se puede observar en la Figura 3 donde aparecen los diagramas de coordenadas paralelas correspondientes a los atributos más destacados. En cada caso se utilizó el color claro para representar un valor bajo en el atributo y el color oscuro para los valores altos.

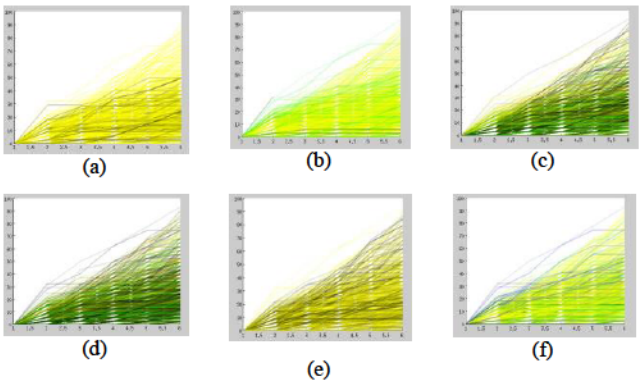

Figura 3. (a) Busca trabajo, (b) Edades, (c) Estudios madre, (d) Estudios padre, (e) Madre trabaja en relación de dependencia, (f) tiempo ingresar facultad

Observando los gráficos de la figura 3 puede decirse que

i) los alumnos que no buscan trabajo (color claro) tienen mejor rendimiento que los que si buscan (Figura 3 (a)).

ii) La edad también condiciona el progreso. Los más jóvenes (color claro) tienden a tener mejor ritmo que los alumnos entre 20-30 años que tienen un ritmo entre medio y bajo, mientras que en los más adultos (color más oscuro) la mayoría tienen un muy bajo rendimiento (Figura 3 (b)).

iii) En cuanto a la actividad laboral de los padres, se observa que los alumnos que tienen madres que no trabajan en relación de dependencia (color claro) tienen un progreso medio o bajo en la carrera (Figura 3 (e)). De forma similar, sucede esto con la actividad de trabajo del padre.

iv) En (Figura 3 (c) y (d)) se observa que a mayor nivel de estudios de los padres (color más oscuro) menor es el progreso del alumno. De forma similar se observa lo mismo con la madre, donde el color oscuro predomina en la parte inferior del gráfico.

v) El tiempo que transcurre entre que egresan del secundario e ingresan a la facultad (Figura 3 (f)), es un factor también relevante. Los que ingresan inmediatamente se distribuyen ampliamente y son los que tienen mejor ritmo (color claro). Los que ingresan dentro de los primeros cinco años (color verde) presentan otro tipo de progreso, que si bien es bueno, no es tal alto como el de los que ingresan casi inmediatamente. Los más oscuros son pocos y muy dispersos.

Resultados

De esta manera, por medio de las visualizaciones, se pudieron encontrar los atributos más representativos de los alumnos regulares y no regulares.

En la Tabla 2 se pueden observar los atributos seleccionados, que constituyen el 17% de la cantidad total.

Ritmo
Busca trabajo
Edad
Tiempo en ingresar facultad
Nivel de estudios Padre
Nivel de estudios Madre
Relación dependencia Madre
Relación dependencia Padre

Tabla 2. Atributos seleccionados

Para medir la efectividad de los atributos seleccionados se construyeron tres tipos de modelos diferentes que permiten clasificar a los alumnos en Regulares y No Regulares. Se utilizaron los siguientes métodos: C4.5, PART y un multiperceptrón entrenado con el algoritmo de backpropagation. [22][23][24]

Para medir el desempeño de cada modelo se utilizó la tasa de acierto y la precisión de cada clase los cuales se calculan de la siguiente forma

$$tasa\_de\_acierto = \frac{t\_pos + t\_neg}{pos + neg} \quad (2)$$

$$precisión(pos) = \frac{t\_pos}{t\_pos + f\_pos} \quad (3)$$

$$precisión(neg) = \frac{t\_neg}{t\_neg + f\_neg} \quad (4)$$

Donde

- *t_pos* y *t_neg* corresponden a la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) correctamente clasificados por el método respectivamente.
- *f_pos* y *f_neg* representan la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) incorrectamente clasificados por el método respectivamente.
- *pos* y *neg* son la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) reales del problema (las respuestas esperadas).

Las Tablas 3 y 4 resumen los resultados obtenidos.

	Sólo atributos Seleccionados	**Todos los atributos**
C4.5	80.1	81.4
PART	80.7	81.1
BPN	80.1	81.2

Tabla 3. Tasa de acierto de cada uno de los modelos.

	Regular		**No regular**	
	(1)	**(2)**	**(1)**	**(2)**
C4.5	86.8	88.9	63.6	63.1
PART	88.6	88.4	61.4	63.5
BPN	86.5	87.9	64.4	64.3

Tabla 4. Precisión de cada clase en cada uno de los modelos. (1) Sólo atributos seleccionados, (2) Todos los atributos

En cuanto a la complejidad del modelo obtenido, la Tabla 5 indica los detalles de cada caso.

	Sólo atributos Seleccionados	**Todos los atributos**
C4.5	7 ramas 13 nodos	41 ramas 81 nodos
PART	12 reglas	230 reglas
BPN	54 neuronas	174 neuronas

Tabla 5. Complejidad de cada modelo

Analizando los resultados obtenidos puede verse en la tabla 3 que la tasa de acierto correspondiente a los atributos seleccionados es ligeramente menor al que se obtiene utilizando todos los atributos. Sin embargo, si se considera que se está trabajando con el 17% de los atributos, este resultado es sumamente favorable.

Por otro lado, si se analiza en la Tabla 4 la precisión para cada clase puede verse que para los métodos C4.5 y BPN la diferencia radica en los alumnos Regulares los cuales constituyen el 70% de la población. Esto hace que las diferencias sean menos significativas. En lo que se refiere a los No Regulares (grupo de interés) el método PART obtiene una precisión superior al 2% utilizando todos los atributos (63.5%) en lugar de utilizar los seleccionados (61.4%). Sin embargo, no

debe dejarse de lado la información de la Tabla 5 donde se detalla la complejidad de cada modelo. Allí se observa que el método PART para obtener la mejora del 2% antes mencionada requiere incrementar considerablemente el número de reglas (en lugar de las 12 reglas con las que consigue una precisión del 61.4% debe utilizar 120 reglas para alcanzar una precisión del 63.5%). En general esto se cumple para todos los modelos ya que operar con un número mayor de atributos incrementa el tamaño del modelo a obtener. Por ejemplo, el método C4.5 obtiene un árbol con una tasa de acierto del 80.1% utilizando 7 hojas y 12 nodos diferentes cuando se genera a partir de los atributos seleccionados. Sin embargo, si se trabaja con el total de los atributos la tasa de acierto mejora sólo en un 1% con un árbol más extenso de 41 hojas y 81 nodos diferentes.

En resumen, observando la Tabla 5 puede afirmarse que utilizando los atributos seleccionados se logra construir un modelo simplificado con una tasa de acierto aceptable.

```
estudios de la madre =< 0.84: Regular

estudios de la madre > 0.84

|   estudios del padre <= 0.2

|   |   relación de dependencia madre=si

|   |   |   relación de dependencia padre=si

|   |   |   |   busca trabajo = no

|   |   |   |   |   edad <=38: No Regular

|   |   |   |   |   edad >38: Regular

|   |   |   |   busca trabajo = si: Regular

|   |   |   relación de dependencia padre=no: Regular

|   |   relación de dependencia madre=no:  Regular

|   estudios del padre > 0.2: Regular
```

Figura 4. Árbol obtenido

A modo de ejemplo, la Figura 4 muestra el resultado de aplicar C4.5 al conjunto de atributos seleccionados.

Calculando la proporción de alumnos que cumplen con cada rama del árbol puede afirmarse que el nivel de estudios de la madre incide directamente en la regularidad de los alumnos. Más del 53% de los alumnos son regulares si la madre tiene a lo sumo estudios terciarios o universitarios incompletos. A mayor nivel de estudios de la madre y el padre con el primario completo, y siendo ambos padres empleados, se encuentran los alumnos no regulares, que cubren el 28% del total. En cambio, si ambos padres tienen estudios secundarios o universitarios los alumnos serán regulares; regla que cubre al 15% de la población.

El árbol descripto en la Figura 4 también puede representarse como un grafo a fin de facilitar su interpretación. Existen herramientas de software libre como RapidMiner que además de construir modelos, permiten crear diferentes visualizaciones de un conjunto de datos disponibles para analizar la relación entre los atributos.

Conclusiones

En este artículo se ha desarrollado un caso de estudio que muestra cómo utilizar la visualización para poder detectar características de un problema en un dominio específico en forma clara y precisa. En este caso particular se pudieron obtener, a partir de la información de los alumnos de la Facultad de Informática de la UNLP, los atributos más representativos para la construcción de un modelo de clasificación que permite describir y caracterizar a los alumnos según su condición de regularidad.

En educación, las herramientas de visualización permiten a los docentes, independientemente del área en la que se desarrollen, hacer uso de las mismas para analizar a sus alumnos debido a que no se necesitan conocimientos específicos de la minería de datos.

En particular, en este trabajo dichas herramientas han permitido obtener un modelo sencillo que posee una tasa de acierto equivalente al construido a partir de la información original.

A futuro se planea incorporar a este análisis la información de los alumnos de la UTN Regional La Plata con el objetivo de establecer similitudes y diferencias entre las poblaciones de alumnos.

Referencias

[1] Koutek, M. Scientific Visualization in Virtual Reality:Interaction Techniques and Application Development. Computer Graphics & CAD/CAM group, Faculty of Information Technology and Systems (ITS), Delft University of Technology (TU Delft), 2003.

[2] Nielson, G. M.; Shriver, B.; Rosenblum, Lawrence. Visualization in Scientific Computing. IEEE Computer Society Press.United States of America, 1979.

[3] Ganuza, M.; Larrea, M.; Martig, S.; Castro, S.; Bjerg, E.; Ferracutti, G. Visualización en Ciencias Geológicas, XIV Workshop de Investigadores en Ciencias de la Computación WICC, 2012.

[4] Charnelli, E. Lanzarini, L. Baldino, G. Diaz, F. Determining the profiles of young people from Buenos Aires with a tendency to pursue computer science studies. XX Congreso Argentino de Ciencias de la Computación CACIC, 2014.

[5] P. Cabena, P. Hadjin ian, R. Stadler, J. Verhees, A.Zanasi, Discovering Data Mining From concept to implementation. Prentice Hall 1997

[6] H. Witten, and E Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, CA: Morgan Kaufmann, 2000.

[7] La Red Martínez, D. L., Acosta, J. C., Cutro, L. A., Uribe, V. E., and Rambo, A. R. Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos. XII Workshop de Investigadores en Ciencias de la Computación. CACIC 2010, pages 162-166.

[8] Luo, Q. Advancing knowledge discovery and data mining. In Knowledge Discovery and Data Mining, WKDD 2008. First International Workshop on.

[9] Valero, S. and Salvador, A. (2009). Predicción de la deserción escolar usando técnicas de minería de datos. In Simposio Internacional en Sistemas Telemáticos y Organizaciones Inteligentes SITOI, 2009.

[10] Rodallegas, E., Torres, A., Gaona, B., Gastelloú, E., Lezama, R., and Valero, S. (2010). Modelo predictivo para la determinación de causas de reprobación mediante minería de datos. In II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA, 2010.

[11] Valero, S., Salvador, A., and García, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. In II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA, 2010.

[12]. Wang, J., Lu, Z., Wu, W., and Li, Y. (2012). The application of data mining technology based on teaching information. In Computer Science Education ICCSE, 2012.

[13] Formia S. Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. Tesis de Especialista en Tecnología Informática aplicada en Educación, 2012.

[14] Asif, R. Merceron, A. Pathan K. Investigating Performances's Progress of Students. CEUR Workshop Proceedings. ISSN 1613-0073, 2014

[15] Bower, A .J. Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis, 2010.

[16] Carpendale, M. S. T. Considering Visual Variables as a Basis for Information Visualization. Technical Report. University of Calgary, Department of Computer Science. 2001.

[17] Larrea, M., Martig, S., Castro. Visualización basada en semántica. XII Workshop de Investigadores en Ciencias de la Computación WICC, 2010.

[18] Sebban, M., Nock, R., Chauchat, J.H., Rakotomalala, R.: Impact of learning setquality and size on decision tree performances. Int. Journal of Computers, Systems and Signals 1, 2000.

[19] Thrun, S.B., Bala et al. The monk's problems a performance comparison of diferent learning algorithms. Technical report, 1991.

[20] Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing ultidimensional geometry. IEEE Visualization, pages 361–378, 1990.

[21] D. K. Urribarri, S. M. Castro, S. R. Martig. Escalabilidad visual en coordenadas paralelas . VIII Workshop de Investigadores en Ciencias de la Computación CACIC, 2006.

[22] Quinlan, R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. 1993

[23] Witten, I. H., & Frank, E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2005

[24] Rosenblatt, Frank. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC. 1961

Dirección de Contacto del Autor/es:
Laura Lanzarini
III-LIDI. Calle 50 y 120 2do. Piso
(1900) La Plata–Prov.de Buenos Aires
Argentina
e-mail: laural@lidi.info.unlp.edu.ar

María Emilia Charnelli
LINTI. Calle 50 y 120 2do. Piso
(1900) La Plata–Prov.de Buenos Aires
Argentina
e-mail: mcharnelli@linti.unlp.edu.ar

Guillermo Baldino
LINSI. Avenida 60 y 124
(1900) La Plata–Prov.de Buenos Aires
Argentina
e-mail: gbaldino@linsi.edu.ar

Javier Díaz
LINTI. Calle 50 y 120 2do. Piso
(1900) La Plata–Prov.de Buenos Aires
Argentina
e-mail: javierd@info.unlp.edu.ar

Laura Lanzarini. Lic. en Informática. Profesora Titular de la UNLP. Investigadora del Instituto de Investigación en Informática III-LIDI, Facultad de Informática. UNLP.

María Emilia Charnelli. Lic. en Informática. Docente de la UNLP. Becaria de Postgrado de la UNLP.

Guillermo Baldino. Ing. en Sistemas de Información. Docente de la UTN. Investigador del Laboratorio de Innovaciones en Sistemas de Información LINSI, Facultad Regional La Plata, UTN.

Javier Díaz. Lic. en Matemática. Profesor Titular de la UNLP. Director del Laboratorio de Investigación de Nuevas Tecnologías Informáticas LINTI, Facultad de Informática, UNLP. Sec. de Relaciones Institucionales de la UNLP.