# Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio.

Sonia Formia[1], Laura Lanzarini[2], Waldo Hasperué[3]

[1] Laboratorio de Informática Aplicada LIA-
Licenciatura en Sistemas-Sede Atlántica, UNRN, Viedma , Argentina
[23] Instituto de Investigación en Informática LIDI,
Facultad de Informática, UNLP, La Plata , Argentina

sformia@unrn.edu.ar;{laural,whasperue}@lidi.info.unlp.edu.ar

## Resumen

En el ámbito de la Universidad Nacional de Río Negro (UNRN), y en particular en la Sede Atlántica desde la Licenciatura en Sistemas, es una creciente preocupación el fenómeno de deserción y desgranamiento que se ha podido apreciar en los cuatro primeros años de vida de la Institución. El presente trabajo describe el proceso de identificación de las características más relevantes del problema a través de las cuales, utilizando técnicas de Minería de Datos (DM), puede obtenerse un modelo de la deserción universitaria en la unidad académica mencionada. Para identificar las características más relevantes se propone analizar, luego del preprocesamiento de los datos, las proyecciones de los atributos en las clases o respuestas esperadas. Su aplicación a los datos de los alumnos de las carreras de grado de la UNRN ha ofrecido resultados satisfactorios permitiendo efectuar recomendaciones tendientes a reducir el porcentaje de alumnos que abandona la carrera.

*Palabras clave*: Selección de atributos. Proyección de atributos. Minería de Datos. Deserción universitaria

## Abstract

At the National University of Río Negro (UNRN), and its Atlantic Coast Delegation in particular, it is an increasing concern for the courses corresponding to the Bachelor's Degree in Systems, the drop-out and crumbling rates observed in the first four years of the Institution. This paper describes the process of identifying the most relevant features of the problem through which, using Data Mining (DM) techniques, a college drop-out model can be obtained for the academic unit mentioned above. In order to identify the most relevant features, after processing the data we will analyze attribute projections for the expected classes or responses. The results of its application to the student data from the courses of the UNRN have been satisfactory, which allows making some recommendations aimed at reducing the percentage of students who drop put from their courses.

*Key words:* attribute selection, attribute projection, data mining, university dropout rates

## 1. Introducción

La organización objeto de estudio es la Universidad Nacional de Río Negro, que habiendo sido creada en el año 2008, comenzó a dictar sus carreras de grado en el año 2009. En la actualidad se dictan un total de 60 carreras de grado. Desde sus inicios ha sido preocupación de las autoridades y de los docentes de las diferentes carreras, el alto índice de deserción y desgranamiento que se observa, a pesar de los pocos años de vida de la Institución. El objetivo principal es poder determinar a priori situaciones potenciales de fracaso académico con el fin de tomar medidas tendientes a minimizar el problema.

En el camino hacia la concreción del objetivo de máxima: predecir la deserción, se pueden encontrar otras metas que aporten información no trivial y de utilidad para la toma de decisiones, por ejemplo, describir o caracterizar a los estudiantes de la UNRN a través de perfiles que ayuden a orientar la implementación de medidas a los estratos en los que las mismas puedan ejercer más influencia positiva.

Distintos autores han propuesto resolver este problema desde distintos enfoques tanto en la captación de estudiantes como en el análisis y detección de abandonos y para la estimación de la duración de la carrera [1] [2] [3] [4] [5] [6] [7]. Recientemente se han desarrollado entornos que facilitan la aplicación de técnicas de DM en contextos educativos [8].

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (*Knowledge Discovery in Databases*) el cual tiene como objetivo la detección automática de patrones existentes en la información disponible sin requerir una hipótesis especificada a priori. 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. Esto último tiene una fuerte incidencia en la técnica a utilizar.

La información disponible proviene del sistema SIU-Guaraní de la UNRN. Es decir que para cada alumno se relevan un número importante de características relacionadas con su situación personal, académica y laboral. Seleccionar de este conjunto las adecuadas para construir un modelo es un verdadero desafío. Existe una relación proporcional entre la cantidad de atributos a utilizar y la complejidad del modelo a obtener. Por lo antes expuesto, resulta de interés la detección temprana de los atributos más relevantes a fin de simplificar el modelo y reducir el tiempo de obtención del mismo.

Este trabajo está organizado de la siguiente forma: la sección 2 describe el preprocesamiento efectuado sobre los datos originales, la sección 3 indica la manera en que fueron considerados, la sección 4 detalla otras estrategias utilizadas previamente para resolver este mismo problema, la sección 5 describe el proceso de selección de atributos a partir de sus proyecciones, la sección 6 ejemplifica un modelo en base a los atributos seleccionados y finalmente la sección 7 expone las conclusiones obtenidas.

# 2. Preparación de datos de la UNRN

Antes de aplicar una técnica de DM específica fue preciso preprocesar los datos a fin de evitar inconsistencias. Esta etapa estuvo guiada por las metodologías de preparación de datos relevadas en la bibliografía y por el conocimiento del dominio.

Se eliminaron atributos con una cantidad excesiva de datos faltantes, se limpiaron valores anómalos, se quitaron atributos constantes y redundantes, se utilizó la generalización para transformar atributos de alta cardinalidad. Se eliminaron atributos no generalizables, se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas, se construyeron nuevos atributos mediante funciones de sumarización (summarize), se discretizaron o numerizaron atributos según la necesidad de los algoritmos y se realizaron normalizaciones de rango. Finalmente, se estableció un atributo de estado que diferencia a los alumnos que ya han abandonado (luego de un año sin actividad académica) de los que cursan normalmente.

# 3. Enfoque del trabajo.

Como primera medida, para comenzar a entender los datos con los que se cuenta y poder describir el dominio del problema, se decidió caracterizar en primer lugar el conjunto de alumnos objeto de estudio, es decir, los registros de alumnos que abandonaron. Motivó esta elección la intención de seleccionar las características relevantes para los alumnos desertores. Entre las tareas descriptivas que provee la minería de datos, el agrupamiento (clustering) es una de las utilizadas con más frecuencia, su objetivo es obtener grupos o conjuntos entre los ejemplos, de manera que los elementos asignados al mismo grupo sean similares [10].

En el caso de estudio la información de la que se dispone incluye datos demográficos, económicos, sociales, familiares y académicos de los alumnos. Mediante la aplicación del algoritmo de agrupamiento k-medias se segmentó a los alumnos desertores en grupos. Las pruebas realizadas pusieron en evidencia la importancia de la dimensionalidad del problema. La gran cantidad de atributos involucrados no permitió que fuera posible encontrar un conjunto de clusters descriptivo de los datos de entrada.

Uno de los problemas centrales en DM es identificar un conjunto representativo de características adecuadas para construir un modelo para una tarea en particular. Los problemas con una alta dimensionalidad, cantidad limitada de ejemplos disponibles y mucha información redundante o irrelevante son difíciles de tratar [9]. El caso de estudio claramente presenta estas características: el SIU-guaraní provee un número importante de atributos para los alumnos y la cantidad de ejemplos se ve limitada por el corto tiempo de vida de la UNRN, disponiéndose tan solo de la información de los últimos 4 años académicos. La vista de la base de datos original (o inicial), utilizada en este trabajo, está formada por 11102 alumnos donde cada uno posee 110 atributos relevados.

En este punto surgió la necesidad de utilizar las herramientas de DM para guiar la selección de un subconjunto de características (atributos) que sean relevantes para el problema.

# 4. Selección de atributos

En [16] utilizando el conjunto de datos de entrada conformado por los grupos de alumnos desertores, se procedió a la transformación del espacio de características mediante dos procesos totalmente diferentes: uno tipo *wrapper* y otro tipo filtro.

Los procesos tipo *wrapper* califican los atributos seleccionados a partir de la performance del modelo que puede construirse a partir de ellos [11]. Existen distintas formas de realizar esto. En [16] se utilizó un proceso de

selección del tipo S*election Forward*. Esta técnica inicia el procedimiento de búsqueda de características evaluando todos los subconjuntos de atributos formados por un solo atributo, luego encuentra el mejor subconjunto de dos atributos, luego de tres y así siguiendo hasta encontrar el mejor subconjunto de características. Para realizar la validación de los conjuntos de características se toma en cuenta la performance de un determinado modelo de aprendizaje. En este caso, se utilizó el método k-medias para agrupar la información disponible. El uso de un algoritmo inductivo posiciona al método dentro de los procesos *wrapper*.

El segundo método de selección implementado está enfocado en la selección genética de características [12]. El algoritmo genético realiza una búsqueda heurística que minimiza el proceso de evolución natural. Para la evaluación utiliza el método CFS (*Correlation-based Features Selection*) que crea un filtro basado en la medida de performance del conjunto de características. Evalúa el valor de un subconjunto de atributos considerando la habilidad predictiva de cada característica junto con el grado de redundancia entre ellas, prefiriendo subconjuntos de atributos altamente correlacionados con la clase que presenten baja intercorrelación entre ellos [13].

Una vez implementados ambos métodos sobre los datos de entrada se pudo apreciar que proporcionaron subconjuntos de atributos similares (ver tabla 1).

A manera de comprobación de la validez de los atributos seleccionados por los métodos descriptos se realizó nuevamente el agrupamiento de los registros pertenecientes a alumnos desertores tomando los atributos seleccionados por el método wrapper y avalado por el resto de los algoritmos presentados. El resultado de la asignación a grupos de esta ejecución se compara con el resultado de la ejecución anterior, determinando que menos de un 10% de los ejemplos se movieron de grupo, lo que indica que el criterio de agrupamiento se conserva a pesar de la reducción de características.

Aplicando luego el mismo agrupamiento a los registros de alumnos no desertores se obtuvo una segmentación comparable en atributos predominantes con la de los alumnos desertores, lo que permite pensar en la utilización de los atributos seleccionados en algoritmos predictivos de abandono.

| Descripción | | Nombre del Atributo | Seleccionado por | |
|---|---|---|---|---|
| | | | Wrapper | Genéti-co |
| El alumno es soltero | | estado_civil = soltero | SI | SI |
| El padre vive | | padre_vive = SI | SI | SI |
| | | situacion_laboral_padre | SI | SI |
| Tiene internet en casa | | alu_tec_int | SI | NO |
| El alumno reconoce que necesita beca | | alu_beca = necesita beca | NO | SI |
| Trabajo actual del alumno | Sit.lab. | alu_trab_sitimp | SI | SI |
| | Relacionado con carrera | rel_trab_carrera | SI | SI |
| | Sueldo Mens. | alu_trab_remmon | SI | SI |
| Sede en la que estudia | | Sede | SI | SI |
| Lugar de nacimiento | | lugar_nacimiento | SI | SI |
| Año egreso secundario | | anio_egreso_sec | NO | SI |
| Piensa trabajar en el futuro | Tipo de trabajo | alu_trab_futtip | SI | SI |
| | horario | alu_trab_futhor | SI | NO |
| Año de nacimiento | | anio_nacim | SI | NO |
| Cant.familiares a cargo | | cant_fami_cargo | SI | SI |
| Cant. hijos del alumno | | cant_hijos_alum | SI | SI |

**Tabla 1**. Lista de atributos seleccionados por los métodos wrapper y genético.

# 5. SOAP: Selección de atributos por proyecciones.

Con el objetivo de reducir aún más la lista de atributos obtenida a través de los métodos descriptos en la sección anterior, se analizó el método SOAP (Selection of attributes by projections) [15].

Este método filtra los atributos utilizando un ranking obtenido de manera determinística reduciendo considerablemente el tiempo de cómputo. SOAP basa su funcionamiento en la incorporación de un nuevo criterio para medir la importancia de un atributo dentro de un marco de aprendizaje supervisado: el número de cambios de etiqueta. Este valor se calcula analizando las proyecciones de los elementos del conjunto de datos sobre cada atributo. De esta manera se pueden ordenar los atributos por orden de importancia en la determinación de la clase.

En SOAP, la selección de atributos se realiza a partir de un único valor: NLC (Number of Label Changes) que relaciona cada atributo con la etiqueta que sirve de clasificación. El NLC se calcula proyectando los ejemplos sobre el eje correspondiente a ese atributo (es decir, ordenando los ejemplos por el atributo), para luego recorrer el eje desde el origen hasta el mayor valor del atributo contabilizando el número de cambios de etiqueta que se producen:

**Algoritmo** SOAP.
**Entrada:** E–conjunto de datos (m ejemplos, n atributos)
**Salida:** R–ranking de atributos

```
R ← {}
para i = 1 to n hacer
   Ordenar E por el atributo Xi
   NLCi ← ContarCambios(E; Xi)
fin para
R (Ranking de atributos según su NLC)
```

**Function** ContarCambios(E; $X_i$)
***Entrada:*** *E–conjunto de datos (m ejemplos, n atributos), $X_i$–atributo a procesar*
***Salida:*** *Cambios–Número de cambios de Etiqueta*

```
R ← {}  Cambios ← 0
para j = 1 to m hacer
 si xj ε Secuencia_Ordenada_Múltiple
  Cambios←Cambios+
          VerCambiosParaMismoValor()
 sino
  si lab(ej) <> lab(ej+1) entonces
        Cambios ← Cambios + 1
     fin si
  fin si
fin para
retornar(Cambios)
```

En ContarCambios es preciso tener en cuenta que al ordenar los valores de un atributo puede ocurrir que aparezcan valores repetidos. Esto dará lugar a una Secuencia_Ordenada_Múltiple (SOM), donde el valor del atributo es el mismo para todos los ejemplos pero las etiquetas asignadas a cada uno de ellos no necesariamente serán iguales. En este caso, se aplica la función VerCambiosParaMismoValor(), la cual calcula el número de cambios correspondientes de la siguiente forma: Si todos los ejemplos poseen la misma etiqueta entonces la cantidad de cambios retornada es cero sino es preciso saber si existe una clase mayoritaria dentro del grupo de ejemplos con el mismo valor del atributo. Si no existe una clase mayoritaria, la cantidad de cambios es la longitud de la SOM menos 1. Si existe una clase mayoritaria, la cantidad de cambios es la cantidad de elementos de la SOM menos la cantidad de elementos de la clase mayoritaria.

## 5.1 Aplicación del algoritmo SOAP al caso de estudio.

Se realizó una implementación del algoritmo descripto anteriormente para aplicarlo al caso de estudio. Se consideraron únicamente los registros correspondientes a los alumnos que abandonaron la carrera agrupándolos previamente mediante k-medias para asignarles etiquetas distintas. Al igual que en los procesos de la sección anterior, se utilizaron 5 grupos (k=5).

Los resultados de la aplicación del algoritmo SOAP a los datos de los alumnos que abandonaron, segmentados en cinco grupos, arrojan un ranking de atributos completo, donde cada atributo recibe un valor de NLC. Los valores del tope del ranking (valores menores de NLC, es decir, atributos que proyectan un menor número de cambios de etiqueta) se presentan en la tabla 2, ordenado por el valor de NLC.

Si se observan los atributos que aparecen en el tope del ranking, se puede ver que predominan los atributos relacionados al trabajo del alumno, junto con los atributos que determinan la edad y las cargas familiares. Estos atributos en general han sido parte integrante de las listas de atributos que se obtuvieron con los primeros algoritmos de selección de características implementados (tabla 1).

Se puede concluir que las características relevantes incluyen principalmente atributos de edad, carga laboral y familiar de los alumnos.

Dada la escasa cantidad de ejemplos con los que se cuenta para la investigación (hecho inapelable por la corta vida de la UNRN) y el gran número de atributos con los que se iniciaron las tareas, se puede inferir que, por el momento, y hasta tanto se pueda contar con más ejemplos para alimentar otros algoritmos, se debe aceptar este grupo de atributos personales y laborales como los

que describen a los alumnos desertores y no desertores. Con esa idea es posible aplicar un algoritmo predictivo, en este caso un árbol de decisión, a la totalidad de los ejemplos disponibles, utilizando un grupo de atributos de los más altos en el ranking del SOAP para clasificar a los alumnos en desertores y no desertores. De esta forma, el árbol que se obtiene es mucho más sencillo que el que se puede alcanzar con el conjunto de atributos original.

| NLC | Atributo |
|---|---|
| 2186 | alu_trab_remmon |
| 2186 | Alu_trab_sitimp |
| 2252 | rel_trab_carrera |
| 3044 | alu_trab_futtip |
| 3571 | alu_trab_futhor |
| 3987 | anio_egreso_sec |
| 4004 | hora_sem_trab_alum |
| 4063 | anio_nacim |
| 5070 | alu_otestsup_uni |
| 5394 | Sit_laboral_madre |
| 5734 | cant_hijos_alum |
| 5747 | Sit_laboral_padre |

**Tabla 2**. Primeros atributos en el ranking según SOAP

| Accuracy :69.84% | True Abandono | True Cursa | Class precision |
|---|---|---|---|
| Predice Abandono | 5174 | 2047 | 71.65% |
| Predice Cursa | 1388 | 2414 | 63.49% |
| Class Recall | 78.85% | 54.11% | |

**Tabla 3**. Performance del árbol construido con los atributos seleccionados (Tabla 2) utilizando el método C4.5 considerando un umbral de confianza de 0.25 y una cantidad mínima de 2 elementos por hoja.

# 6. Modelo predictivo de deserción.

Se realizaron pruebas con el algoritmo C4.5 [14] utilizando los nueve primeros atributos del ranking SOAP (tabla 2), exceptuando "anio_egreso_sec" que corresponde al año en que el alumno egresó del colegio secundario. Esto último tiene que ver con que el atributo "anio_nacim", cuyo valor representa el año de nacimiento del alumno, ya está incluido y que ambos atributos están estrechamente correlacionados. La tabla 3 muestra la performance del modelo obtenido.

Puede observarse que la lista de atributos utilizada tiene una longitud equivalente al 56.25% de la lista de atributos de la tabla 1 (9/16 = .5625). Los atributos seleccionados permiten construir un modelo capaz de predecir correctamente el 71.65% de los casos de abandono. La tasa de acierto es menor en el caso de tener que predecir si el alumno sigue cursando.

La Figura 1 muestra una versión podada del árbol resultante considerando un número mínimo de 10 elementos por hoja.

```
anio_nacim <= 1989: Abandono (7614/2538)
anio_nacim > 1989
|  anio_nacim <= 1992
|  |  alu_otestsup_uni = S
|  |  |  Alu_trab_sitimp = Relación de dependencia: Cursa (65/21)
|  |  |  Alu_trab_sitimp = no trabaja: Cursa (257/90)
|  |  |  Alu_trab_sitimp = Monotributista: Abandono (11/4)
|  |  alu_otestsup_uni = N
|  |  |  Alu_trab_futtip = Obrero o empleado (asalariado)
|  |  |  |  Alu_trab_futhor = Más de 10 y hasta 20 hs: Cursa (93/42)
|  |  |  |  Alu_trab_futhor = No trabajaré: Cursa (7/2)
|  |  |  |  Alu_trab_futhor = 35 o más horas
|  |  |  |  |  Hora_sem_trab_alum = No informado: Abandono (33/13)
|  |  |  |  |  Hora_sem_trab_alum = de 21 a 35 hs.: Abandono (0)
|  |  |  |  |  Hora_sem_trab_alum = hasta 20 hs.: Cursa (4/1)
|  |  |  |  |  Hora_sem_trab_alum = de 36 o más hs.: Cursa (1)
|  |  |  |  |  Hora_sem_trab_alum = no trabaja: Abandono (48/17)
|  |  |  |  Alu_trab_futhor = Hasta 10 horas: Abandono (128/51)
|  |  |  |  Alu_trab_futhor = Más de 20 y menos de 35 horas
|  |  |  |  |  Hora_sem_trab_alum = No informado: Abandono (25/10)
|  |  |  |  |  Hora_sem_trab_alum = de 21 a 35 hs.: Cursa (3/1)
|  |  |  |  |  Hora_sem_trab_alum = hasta 20 hs.: Abandono (2/1)
|  |  |  |  |  Hora_sem_trab_alum = de 36 o más hs: Abandono (36/15)
|  |  |  |  |  Hora_sem_trab_alum = no trabaja: Cursa (11/3)
|  |  |  Alu_trab_futtip = No trabajaré: Cursa (1358/613)
|  |  |  Alu_trab_futtip = Cuenta propia
|  |  |  |  Alu_trab_remmon = de 1200 a 2000$: Abandono (9/4)
|  |  |  |  Alu_trab_remmon = más de 3000$: Cursa (2)
|  |  |  |  Alu_trab_remmon = no trabaja: Abandono (35/16)
|  |  |  |  Alu_trab_remmon = hasta 1200$: Cursa (33/14)
|  |  |  |  Alu_trab_remmon = de 2000 a 3000$: Abandono (2/1)
|  |  |  Alu_trab_futtip = Desconoce
|  |  |  |  Alu_trab_futhor=Más de 10 y hasta 20 horas:Abandono (81/28)
|  |  |  |  Alu_trab_futhor = No trabajaré: Abandono (107/52)
|  |  |  |  Alu_trab_futhor = 35 o más horas: Abandono (6/3)
|  |  |  |  Alu_trab_futhor = Hasta 10 horas: Cursa (138/66)
|  |  |  |  Alu_trab_futhor= Más de 20 y menos de 35 hs: Abandono (29/9)
|  anio_nacim > 1992
|  |  Hora_sem_trab_alum = No informado: Cursa (827/278)
|  |  Hora_sem_trab_alum = de 21 a 35 hs.: Cursa (13/5)
|  |  Hora_sem_trab_alum = hasta 20 hs.: Cursa (26/9)
|  |  Hora_sem_trab_alum = de 36 o más hs.: Abandono (10/4)
|  |  Hora_sem_trab_alum = no trabaja: Abandono (9/2)
```

**Figura 1.** Árbol de decisión para determinar deserción

## Conclusiones

Este trabajo presenta la aplicación de un método de selección de características basado en proyecciones que es capaz de operar sobre atributos nominales y numéricos de manera supervisada. A partir del ranking que establece entre los atributos es posible determinar un punto de corte para identificar los más representativos. En este caso, su aplicación permitió reducir la lista original (tabla 1) en más de un 40% (los 9 primeros atributos de la tabla 2).

La tabla 2 muestra que los atributos más relevantes son los relacionados con la situación laboral del alumno tanto en lo que se refiere a su trabajo actual como a sus intenciones de trabajar en el futuro.

Como producto preliminar se pueden obtener orientaciones claras para guiar las acciones a tomar a favor de la disminución de la deserción en la UNRN: es claro que las variables laborales de los alumnos tienen marcada influencia en su posibilidad de permanencia en los claustros, de manera que acciones directas sobre esta realidad, como el aumento de las becas otorgadas, podría brindar un camino a seguir.

Más allá de estas conclusiones, se dejó planteado un modelo predictivo que puede ser mejorado a lo largo del tiempo y con la incorporación de más ejemplos al conjunto de datos.

## Referencias


[1]. La Red Martínez, D. L., Acosta, J. C., Cutro, L. A., Uribe, V. E., and Rambo, A. R. (2009). Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos. In XII Workshop de Investigadores en Ciencias de la Computación – CACIC 2010, pages 162–166.

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

[3]. Alcover, R., Benlloch, J., Blesa, P., Calduch, M. A., Celma, M., Ferri, C., Hernández Orallo, J., Iniesta, L.,Más, J., Ramírez Quintana, M. J., Robles, A., Valiente, J. M., Vicent, M. J., and Zúnica, L. R. (2007). Análisis del rendimiento académico en los estudios de informática de la universidad politécnica de valencia aplicando técnicas de minería de datos. Technical report, Universidad Politécnica de Valencia.

[4]. 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, pages 332–340.

[5]. 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, pages 48–55.

[6]. 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, pages 33–39.

[7]. 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 7th International Conference on, pages 652 –657.

[8]. Ngo, L., Dantuluri, V., Stealey, M., Ahalt, S., and Apon, A. (2012). An architecture for mining and visualization of u.s. higher educational data. In Proceedings of the 2012 Ninth International Conference on Information Technology - New Generations, ITNG '12, pages 783–789, Washington, DC, USA. IEEE Computer Society.

[9]. Hernández Orallo, J., Ramírez Quintana, M., and Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. Ed. Pearson.

[10]. Witten, I. H. and Frank, E. (2011). Data Mining: Practical Machine Learning.Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco, CA, 3th edition.

[11]. Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artif. Intell., 97(1-2):273–324.

[12]. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st ed.

[13]. Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand.

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

[15]. Sanchez, Roberto Ruiz. Heurísticas de selección de atributos para datos de gran dimensionalidad. Tesis doctoral Universidad de Sevilla. 2006

[16]. Formia S, Lanzarini L. Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. VIII Congreso de Tecnología en Educación y Educación en Tecnología 2013. Sgo.del Estero. Junio 2013.

*Dirección de Contacto del Autor/es*:
Sonia Formia
Patriarca 388
(8500) Viedma- Río Negro
Argentina
e-mail: sformia@unrn.edu.ar

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

Waldo Hasperué
III-LIDI. Calle 50 y 120 2do. Piso
(1900) La Plata – Prov.de Buenos Aires
Argentina
e-mail: whasperue@lidi.info.unlp.edu.ar

**Sonia Formia**. Especialista en Tecnología Informática aplicada en Educación (UNLP). Ingeniera en Sistemas (UNICEN). Profesora Adjunta Licenciatura en Sistemas UNRN. Investigadora Laboratorio de Informática Aplicada – Sede Atlántica UNRN.

**Laura Lanzarini.** Lic. en Informática. Profesora Titular de la UNLP. Investigadora del Instituto de Investigación en Informáctica LIDI – Facultad de Informática. UNLP

**Waldo Hasperué.** Dr. en Informática. Jefe de Trabajos Prácticos. Investigador del Instituto de Investigación en Informáctica LIDI – Facultad de Informática. UNLP