# Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información

**Horacio Kuna, Ramón García Martínez, Francisco R. Villatoro**
Dept. Informática, Facultad de Ciencias Exactas Químicas y Naturales, U.N.Misiones
Área Ingeniería del Software. Lic. en Sistemas. Dept. Desarrollo Productivo y Tecnológico. UNLa
Laboratorio de Sistemas Inteligentes, Facultad de Ingeniería.UBA.
Dept. Lenguajes y Ciencias de la Computación, Universidad de Málaga.
hdkuna@unam.edu.ar,rgarciamar@fi.uba.ar,villa@lcc.uma.es

## Resumen

El abandono de los estudios universitarios en el nivel de pregrado, es un fenómeno global en el Sistema Universitario Argentino, que conlleva la necesidad de desarrollar políticas de retención de estudiantes. Estas políticas requieren la identificación de las posibles causas de deserción. En este articulo se presenta el uso de algoritmos TDIDT para descubrir reglas que caractericen el abandono a partir de la información disponible en el Sistema SIU-Guarani. El abordaje empleado ha permitido identificar las variables con mayor incidencia en la deserción.

**Palabras claves**: Deserción, identificación de Causas de Deserción, Algoritmos TDIDT.

## 1. Introducción

En la mayoría de las Universidades públicas de la Argentina, el ingreso es irrestricto para las carreras de grado y pregrado, las últimas cifras oficiales [SPU, 2009] correspondientes al año 2006, indican que existen alrededor de 1.300.000 de estudiantes universitarios, ingresando por año al sistema alrededor de 270.000 alumnos y egresando en el año 2005 aproximadamente 64.000 profesionales.
Se verifica un fenómeno muy preocupante a nivel global que es la deserción. Se puede definir a la deserción como el abandono por parte de un alumno de los estudios formales de una determinada carrera [Parrino, 2004].
Un estudiante que abandona una carrera puede seguir muchos caminos, uno de ellos puede ser continuar estudiando otra carrera de la misma Facultad o Universidad, puede cambiar de Universidad, puede continuar los estudios años después o definitivamente no volver a pisar los claustros universitarios y comenzar a trabajar o convertirse en desocupados.
Para algunos autores [Manski, 1989], el abandono de una carrera por parte de un alumno no necesariamente es malo, ya que por ejemplo el paso de una persona por las aulas de una Universidad pudo significar un crecimiento personal, pudo agregar conocimientos útiles y aplicables en su vida, entre otros.
El propio concepto de deserción es muy discutido, pero lo que es claro es que para el estado implica un enorme costo, para el estudiante puede significar algún tipo de frustración y requiere de políticas de estado que la prevengan. El fenómeno del degranamiento es particularmente relevante en el primer año de estudios, la figura 1 muestra la deserción por cohorte en pregrado en algunos paises Latinoamericanos y del Caribe [IESALC-UNESCO, 2005].

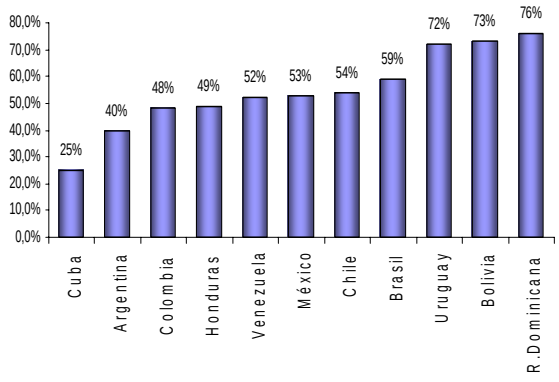



**Fig. 1**. Deserción por cohortes en pregrado

Muchos autores presentan distintas hipótesis que tratan de explicar los fenómenos de la deserción y retención de los estudiantes:

- Algunos explican estos fenómenos a partir de dos teorías sociológicas: "El modelo de integración del estudiante" [Spady, 1970; Tinto, 1975] donde la integración del estudiante al mundo académico afecta en forma directa a la determinación de abandonar o no los estudios, otro es el "Modelo de desgaste del estudiante" [Bean, 1980] que da relevancia a los factores externos a la institución educativa.
- Para [Hackman *et al*, 1970] el problema fundamental de la deserción tiene que ver con la ausencia de interés y no con la imposibilidad por parte del alumno de cumplir con los requisitos que la Universidad exige.
- Según [Braxton *et al,* 1997] de acuerdo a la relevancia que se le otorga a las variables que intentan explicar el fenómeno de la deserción y retención, sean familiares, individuales o institucionales, aborda distintas dimensiones de análisis: Psicológicos, Económicos, Sociológicos, Organizacionales y de Interacción.
- Para [Aparicio *et al*, 2006] centra las causales del abandono en temas relacionados con procesos vocacionales.

Existen una variada cantidad de interpretaciones que intentan explicar el fenómeno de la deserción, al tratarse de una compleja problemática ninguna de estas interpretaciones agota el tema.

El Ministerio de Educación de la República Argentina, a través de la Secretaría de Políticas Universitarias (SPU) tiene una profunda preocupación sobre la problemática de la deserción, en el octubre del año 2008 organizó un Seminario Internacional sobre la Deserción, donde se analizó en profundidad la problemática.

Complementando las distintas hipótesis existentes sustentadas en algunos casos en métodos estadísticos, el Consorcio SIU [2009] conformado por 33 Universidades Nacionales de la Argentina, realiza distintos estudios que no parten de hipótesis predeterminada sino que intentan encontrar patrones de comportamiento en forma automática en bases de datos de los sistemas de gestión académica de las Universidades.

En particular el trabajo que se presenta se relaciona con la Universidad Nacional de Misiones, la misma se asienta en una provincia que tiene particulares características dentro del contexto nacional ya que el 90% de sus fronteras son de carácter internacional, encontrándose en el corazón del MERCOSUR, su economía se sustenta en cultivos regionales, como la yerba mate, te, entre otros, la explotación forestal y la industria del turismo. Posee alrededor de 1.000.000 de habitantes con una fuerte inmigración proveniente de Europa central. Su Universidad cuenta con alrededor de 20.000 estudiantes, ingresando en el año 2006 aproximadamente 3.600 alumnos y egresando en el año 2005 cerca de 600 profesionales. El estudio que se presenta en este trabajo se realizó en una Facultad con carreras técnicas.

## 2. Explotación de Información

En [Britos, 2008] se sostiene que explotación de información se ha definido como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información. Esto resulta una alternativa de solución a problemas que no pueden ser resueltos mediante algoritmos tradicionales, entre los cuales podemos mencionar especificación de condiciones asociadas a diagnósticos técnicos o clínicos, identificación de características que permitan reconocimiento visual de objetos, descubrimiento de patrones o regularidades en estructuras de información (en particular en bases de datos de gran tamaño), entre otros.

Los métodos tradicionales de análisis de datos incluyen el trabajo con variables estadísticas, varianza, desviación estándar, covarianza y correlación entre los atributos; análisis de componentes (determinación de combinaciones lineales ortogonales que maximizan una varianza determinada), análisis

de factores (determinación de grupos correlacionados de atributos), análisis de clusters (determinación de grupos de conceptos que están cercanos según una función de distancia dada), análisis de regresión (búsqueda de los coeficientes de una ecuación de los puntos dados como datos), análisis multivariable de la varianza, y análisis de los discriminantes. Todos estos métodos están orientados numéricamente. Son esencialmente cuantitativos.

En contraposición, los métodos de la explotación de información permiten obtener resultados de análisis de la masa de información que los métodos convencionales no logran tales como: los algoritmos TDIDT, los mapas auto organizados (SOM) y las redes bayesianas. Los algoritmos TDIDT permiten el desarrollo de descripciones simbólicas de los datos para diferenciar entre distintas clases [Quinlan, 1986; 1990]. Los mapas auto organizados pueden ser aplicados a la construcción de particiones de grandes masas de información. Tienen la ventaja de ser tolerantes al ruido y la capacidad de extender la generalización al momento de necesitar manipular datos nuevos [Kohonen, 1982; 1995]. Las redes bayesianas pueden ser aplicadas para identificar atributos discriminantes en grandes masas de información, detectar patrones de comportamiento en análisis de series temporales. [Heckerman *et al*., 1995].

## 3. Diseño experimental y Variables

El objetivo del trabajo fue tratar de encontrar en forma automática las posibles causas de abandono entre el primer y segundo año de una Facultad con carreras Técnicas (variable independiente). Se trató de obtener patrones automáticos de comportamiento de la base de datos del sistema de gestión academico SIU-Guaraní mediante el uso de procesos de explotación de información estandarizados [Britos, 2008] .

Se trabajó con los siguientes datos (variables dependientes):

- datos propios del alumno (por ejemplo: edad al momento de ingreso, colegio secundario en que estudio, entre otros.),
- datos referentes al rendimiento académico en la carrera durante el primer año
- datos propios del alumno que se modifican con el tiempo (estado civil, si tiene hijos, si trabaja, situación económica, entre otros.).

Se realizó un preprocesamiento con el objetivo de mejorar la calidad de los datos y se detectaron algunos problemas relacionados con datos faltantes, incompletos e inconsistentes que fueron depurados para optimizar el proceso de explotación de información.

Algunas variables fueron descartadas ya que no brindaban información sustantiva al objetivo planteado y fueron creadas nuevas variables a partir de variables ya existentes.

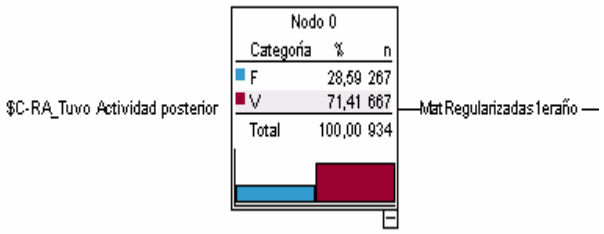
Se tomo como punto de corte lo acontecido durante el año de ingreso a la carrera y lo que sucede luego de transcurrido el primer año del alumno, Se considera tener actividad posterior al primer año como realizar alguna de estas actividades: rendir algún final, cursar alguna materia (más allá de aprobarla o no) u obtener alguna equivalencia. No fueron considerados los postgrados, solamente las carreras de grado y pregrado. Para el estudio realizado se tomaron los datos del año 2003. Las principales variables utilizadas se muestran en la Tabla 1 (al final del trabajo).

## 4. Resultados e Interpretación

### 4.1 Resultados

El principal objetivo fue encontrar características de alumnos que permitan explicar algunas de las causales que pueden generar que un alumno deje de tener actividad académica después del primer año de cursado. La variable objetivo planteada fue *tuvo_actividad_posterior,* definiéndose como actividad posterior durante el segundo año: cursar y/o rendir una materia u obtener una equivalencia.

Para llegar al objetivo propuesto se aplicó un algoritmo de inducción que permitió obtener árboles de decisión y a partir de los mismos reglas que posibilitan explicar porque los alumnos no tuvieron actividad durante el segundo año de cursado. El conocimiento que surgió llamo la atención a los expertos del dominio, en particular en las siguientes reglas obtenidas a partir del nodo raíz:




El nodo raíz permite visualizar que el 28,59% de los alumnos que ingresaron durante el año académico 2002 no tuvieron actividad académica durante el año 2003. Algunas de las reglas obtenidas son:

Regla 1:

- *mat_regularizadas_primer_año* <=1 (62,38%)
  - *costea_estudios* = Con su trabajo (58,82%)
    - *tipo_titulo_secundario* = Bachiller (83,33%) then *tuvo_actividad_posterior* = F (28,59%)

Esta regla brinda conocimiento sobre la principal causa por la cual un alumno no tuvo actividad durante el segundo año, y es la cantidad de materias regularizadas en el primer año, el 62,38% de los alumnos que regularizaron 1 o ninguna materia no tuvieron actividad posterior en el segundo año, de este universo el 58,82% que costeo sus estudios con su trabajo abandono al segundo año de carrera y dentro de este grupo es importante el peso de la variable "*tipo_titulo_secundario*" ya que el 83,33% que tiene el titulo de Bachiller no tuvo actividad en el segundo año.
Esta regla ha llamado la atención de los expertos en el dominio y ha permitido identificar como alumnos potencialmente en riesgo de abandonar los estudios en el segundo año a aquellos que regularizaron 1 o ninguna materia, costean sus estudios con su trabajo y tienen titulo de Bachiller.

Regla 2:

- *mat_regularizadas_primer_año* <=1 (62,38%)
  - *costea_estudios* = con el aporte de familiares u otros (31,01%)
    - *finales_des_y_ausentes_primer_año* <=3 (35,14%)
      - *dif_egreso_sec__ingreso* = Entre 8 y 15 años (44,44%)
        - *viaja* = Si (100%) then *tuvo_actividad_posterior* = F (28,59%)

Esta regla brinda conocimiento sumamente útil a los expertos en el dominio al aparecer las variables "*finales_des_y_ausentes_primer_año*","*dif_egreso_sec__ingreso*" y *viaja*, junto a las dos variables que mayor peso tienen en el abandono de los alumnos en el segundo año, detectadas en la regla anterior (cantidad de materias regularizadas y forma en que costea sus estudios).
Aparece entonces un nuevo grupo de alumnos en riesgo de abandonar sus estudios que son aquellos que regularizaron 1 o ninguna materia (el 62,38% no tuvo actividad en el segundo año), dentro de ese universo el 31,01% que costea sus estudios con el aporte de familiares no continuó sus estudios, dentro de ese grupo el 35,14% que tuvo 3 o menos de 3 finales ausentes o desaprobados no continuó con sus estudios en el segundo año, la aparición de esta variable llamó poderosamente la atención de los expertos, dentro de ese universo mas del 44% de alumnos que pasaron entre 8 y 15 años desde que egresaron del secundario hasta el ingreso a la universidad no continuó en el segundo año y dentro de ese universo el 100% debía viajar más de 10 kilómetros desde su residencia hasta la facultad.

Regla 3:

- *mat_regularizadas_primer_año* >1 (88,94%)
  - *cursados_des_y_ausentes_primer_año* <= 1 (96,47%) then *tuvo_actividad_posterior* = V (71,41%)

Esta regla brinda conocimiento para el caso de la permanencia de los alumnos en el segundo año y el peso que tiene tener mas de una materia regularizada en el primer año (el 88,94% tuvo actividad en el segundo año de la carrera) y dentro de ese universo el 96,47% tuvo menos de una materia cursada desaprobada en el primer año.

### 4.2 Interpretación

El conocimiento que surge en la base de datos del sistema de gestión de alumnos, permite detectar como un denominador común en los alumnos que abandonan al segundo año el no regularizar materias en el primer año de cursada, como contraparte la mayoría de alumnos que regulariza mas de una materia continúa sus estudios, esto implica que la permanencia depende directamente de esta variable.

Ahora bien, que particularidad tienen los alumnos que no regularizan materias y que abandonan al segundo año, el proceso de explotación de información muestra que las variables que representan la forma en que costean sus estudios, el tipo de titulo secundario, los años que pasaron desde el egreso del secundario, si el alumno debe viajar más de 10 km desde su residencia hasta la facultad, tienen un peso significativo en la población que abandona sus estudios.

Muchas de estas variables aparecen como hipótesis de causas de abandono para muchos autores, lo particular de este trabajo tiene que ver con la detección de la cantidad de materias regularizadas, este es el primer indicador temprano que se debe observar para establecer políticas que permitan lograr la permanencia.

## 5. Conclusiones y futuras líneas de Investigación

Analizando los resultados obtenidos después del proceso de explotación de la información aplicando algoritmos de inducción, es posible afirmar que estas herramientas resultan de gran importancia para determinar las causales de la deserción, dando elementos para realizar una análisis institucional y tomar decisiones que permita definir nuevas estrategias dentro de la institución, como por ejemplo orientar la política de becas, cursos de nivelación para egresados de escuelas secundarias no afines a la carrera Universitaria, clase de apoyo para disminuir los desaprobados, entre otros. Es importante destacar que la confiabilidad de los resultados del proceso de explotación de información tiene directa relación con la calidad de los datos de los sistemas de gestión.

Como consecuencia de estas conclusiones surgen una serie de preguntas con relación a los datos que se recogen de cada persona que se inscribe en una carrera: ¿son los necesarios? ¿Son pocos? ¿Son bien interpretados? ¿Son excesivos? ¿Están bien categorizados? ¿Se necesita incorporar datos nuevos? ¿Se debe realizar un control de calidad mas exhaustivo de los datos que están en la base de datos?

## 6. Bibliografía


Aparicio, M. Garzuzi, V (2006): *Dinámicas Identitarias, Procesos Vocacionales y su Relación con el Abandono de los Estudios. Un Análisis en Alumnos Ingresantes a la Universidad.* Revista De Orientación Educacional V20. Pp 15-36

Bean, J. P. (1980): *Student Attrition, Intensions and Confidence*. In: Research in Higher Education 17. Pp 291-320.

Britos, P. (2008). *Procesos de Explotación de Información Basados en Sistemas Inteligentes*. Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática de la Universidad Nacional de La Plata. http://laboratorios.fi.uba.ar/lsi/td-pb-fi-unlp.pdf. Pagina vigente al 4/05/09.

Hackman, J. y Dysinger, W. S. (1970): *Commitment To College as a Factor in Student Attrition.* Sociology of Education, 1970, 43 (3), 311-324.

Heckerman, D., Chickering, M., Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data*. Machine learning 20: 197-243.

IESALC-UNESCO (2005). *Datos para Colombia*. SNIES, Ministerio de Educación Nacional. Colombia.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag Publishers.

Mansky, C. (1989): *Anatomy of The Selection Problem. Journal of Human Resources* 24, 343-360.

Parrino, M. (2004): *De la Reflexión a la Acción Política para Disminuir los Procesos de Deserción Universitaria*. IV Coloquio Internacional sobre Gestión Universitaria en America Del Sud. Floreanopolis.

Quinlan, J. (1986). *Induction of decision trees*. Machine Learning, 1(1): 81-106.

Quinlan, J. (1990). *Learning Logic Definitions from Relations.* Machine Learning, 5:239-266

Kohonen, T. (1982). *Self-organized formation of topologically correct feature maps*. Biological Cybernetics, 43: 59-69.

SIU (2009). *Sistema Inter Unversitario*. http://www.siu.edu.ar/. Pagina vigente al 4/05/09.

Spady, W. (1970): *Dropouts From Higher Education: An Interdisciplinary Review And Synthesis*. Intechange 1. Pp.64-85.

SPU (2009). *Secretaria de Políticas Universitarias*. Ministerio de Educación. Argentina. http://www.me.gov.ar/spu/. Pagina vigente al 4/05/09.

Tinto (1975). *Dropout From Higher Education: A Theoretical Synthesis ff Recent Research.* Review of Educational Research 45. Pp. 89-125.

| Nombre de la variable | Tipo de variable | Descripción | Valores posibles |
|---|---|---|---|
| *costea_estudios* | Dependiente | Forma en la que costea sus estudios un alumno | Con el aporte de familiares u otros.<br>Con su trabajo.<br>Con su trabajo y el aporte de familiares. |
| *cursados_des_y_ausentes_primer_año* | Dependiente | Cantidad de materias cursadas y desaprobadas o ausentes durante el primer año | <=1<br>>1 |
| *dif_egreso_sec__ingreso* | Dependiente | Cantidad de años que pasaron desde que el alumno finalizó la secundaria e ingresó a la facultad | 16 o mas años.<br>Entre 3 y 7 años.<br>Entre 8 y 15 años.<br>Menos de 3 años. |
| *Finales_des_y_ausentes_primer_año* | Dependiente | Cantidad de finales desaprobados y ausentes durante el primer año | <=3<br>>3 |
| *mat_regularizadas_primer_año* | Dependiente | Cantidad de materias regularizadas durante e primer año de estudios | 1 a n |
| *tipo_titulo_secundario* | Dependiente | Titulo secundario obtenido por el alumno | Bachiller.<br>Educ.polimodal.<br>Maestro mayor de obras.<br>Perito mercantil.<br>Técnico<br>Otros |
| *Tuvo_actividad_posterior* | Independiente | *Materias cursadas* + *materias rendidas* + *materias aprobadas por equivalencia* en el segundo año de cursado | V = Verdadero<br>F = Falso |
| *Viaja* | Dependiente | Define si el alumno debe viajar mas de 10 km desde su lugar de residencia y la facultad. | Si.<br>No. |

**Tabla 1.** Principales variables utilizadas en el proceso de explotacion de información