Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios

Autores
Ing. Sonia Alejandra Formia
Licenciatura en Sistemas – Sede Atlántica – UNRN
e-mail: sformia@unrn.edu.ar

Lic. Laura Lanzarini
Facultad de Informática - UNLP
e-mail: laural@lidi.info.unlp.edu.ar

Resumen

En el ámbito de la Universidad Nacional de Río Negro (UNRN), y en particular desde la Licenciatura en Sistemas, es una creciente preocupación del cuerpo docente el fenómeno de deserción y desgranamiento que se ha podido apreciar en los tres primeros años de vida de la carrera.
El presente trabajo tiene como objetivo brindar una breve descripción de la tarea realizada para abordar el estudio del fenómeno de deserción estudiantil universitaria mediante un proceso de extracción de conocimiento a partir de datos. La etapa más relevante de este proceso es la minería de datos, que provee mecanismos para la extracción no trivial de información implícita, previamente desconocida a partir de una base de datos y así descubrir reglas y/o patrones significativos de información que puedan ayudar tanto en el diagnóstico correcto del problema como en la formulación de estrategias de solución.
Se realizó una evaluación de las diferentes técnicas y su posibilidad de aplicación en el caso de estudio, poniendo énfasis en los métodos no supervisados. Luego se realizaron pruebas de concepto con algunas de las técnicas que resultaron de interés utilizando las bases de datos de los alumnos de las carreras de grado de la UNRN. Finalmente se describen los resultados obtenidos.

Palabras clave: deserción universitaria – minería de datos – extracción de conocimiento

1. Motivación y objetivo

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 consta de cuatro sedes (Andina, Alto Valle, Valle Medio y Atlántica) en las que 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.

2. Extracción de Conocimiento

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (*Knowledge Discovery in Databases*) el cual consta de una serie de fases que definen la metodología a utilizar. La secuencia de estas fases no es estricta y

frecuentemente hay movimiento entre ellas, dependiendo del resultado de cada fase dando lugar a un proceso de naturaleza cíclica (3).

Fase 1. Comprensión del Dominio.

Esta fase involucra pasos clave como determinar los objetivos, comprender la situación, determinar el papel del DM en el proyecto y visualizar un plan de trabajo.

Fase 2. Recopilación e integración de datos.

Esta fase se inicia con la obtención de los datos, se procede a familiarizarse con ellos e identificar su procedencia. En esta etapa se trabaja en recolectar los datos, describirlos, explorarlos y verificar su calidad.

Fase 3. Preparación de los datos.

Es necesario seleccionar y preparar el subconjunto de datos a minar, denominado vista *minable*. Esta fase cubre todas las actividades para construir el conjunto final de los datos que serán utilizados en las herramientas de modelado, incluye la selección de archivos, registros y atributos, así como la transformación y limpieza de los datos.

Fase 4. Modelado.

También denominada Minería de Datos, por ser la más característica del KDD, es la fase en la que se seleccionan y aplican diferentes técnicas de modelado, configurando sus parámetros para la obtención de resultados. Aquí es donde se produce conocimiento nuevo, construyendo modelos a partir de los datos recopilados.

Fase 5. Interpretación y evaluación.

Los modelos obtenidos en la fase anterior son interpretados y evaluados a fin de comprobar si cumplen los objetivos planteados en las fases preliminares. Aquí es crítico determinar si partes importantes de la realidad han sido lo suficientemente consideradas y se debe decidir sobre la utilización de los resultados del proceso de DM.

Fase 6. Difusión y uso de los resultados.

La creación del modelo no implica la finalización del proyecto. El conocimiento obtenido debe ser organizado y presentado de manera que pueda ser comprendido y utilizado por el usuario final.

La tarea importante de esta fase consiste en que el usuario entienda los resultados y pueda utilizar los modelos creados.

Fase 7. Implementación de medidas basadas en el conocimiento obtenido.

Cuando la fase de uso de resultados genera una clase de conocimiento que habilita al usuario a ejecutar acciones en pos de resolver el problema planteado originalmente, se produce una etapa de implementación de medidas que debe llevar a cabo la organización. Estas medidas tendrán como objetivo mejorar o corregir la realidad descubierta a través del modelado, actuando directamente sobre la organización.

Fase 8. Medición de resultados.

Luego de la implementación de las medidas de la fase anterior, es posible la utilización del DM para medir los resultados alcanzados por esas acciones En esta fase se pueden volver a ejecutar los modelos para compararlos con los obtenidos en la primera iteración y de esa manera conseguir mediciones concretas del éxito o fracaso de las medidas tomadas.

3. Preparación de datos de la UNRN

La UNRN utiliza el sistema SIU-Guaraní para la gestión académica; el mismo almacena los datos en una base de datos relacional. El relevamiento realizado con los responsables de la administración de los datos del Rectorado de UNRN dejó en claro que las tablas del modelo de datos del SIU-Guaraní son la fuente principal de datos de la Universidad en lo que a alumnos se refiere, y que por el momento no se utilizan fuentes externas que provean otra información relevante. La recopilación de datos se reduce entonces a la obtención de la definición de todas las tablas del modelo relacional y los datos de cada una de ellas.

La preparación de datos, en general, organiza y representa las *vistas minables* a las que se les aplicarán las técnicas concretas de minería de datos. Esta organización de los datos debe ir acompañada de una limpieza e integración de los mismos para que estén en condiciones de ser analizados. Además, debido a las características propias de las técnicas de minería de datos, es necesario generalmente

hacer transformaciones en los datos antes de utilizarlos.

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 exceso de datos faltantes, se limpiaron valores anómalos (producto de errores de carga), se eliminaron atributos constantes (como la unidad académica, con valor 'UNRN' para toda la base de datos) y redundantes (como las claves: número de inscripción y legajo), se utilizó la generalización para transformar atributos de alta cardinalidad como el colegio secundario que fue reemplazado por el tipo de colegio ('Estatal' o 'Privado'). Se eliminaron atributos no generalizables como el domicilio del estudiante. También se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas. Por ejemplo, la localidad de nacimiento fue reemplazada por la zona geográfica correspondiente.

Por otro lado se construyeron nuevos atributos mediante funciones de sumarización, agregando la cantidad de materias aprobadas, desaprobadas, promocionadas y calculando promedios de notas finales.

Una tarea más de preparación de datos, relacionada directamente con los requerimientos de algunos algoritmos de minería de datos utilizados, fue la discretización o numerización, según corresponda, de algunos atributos y la posterior normalización de rango (en particular para el uso de algoritmos basados en distancias). También 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.

Selección de atributos o características.

Uno de los problemas centrales en la Minería de Datos es identificar un conjunto representativo de características adecuadas para construir un modelo para una tarea en particular. Hay muchos factores que afectan el éxito de una tarea de DM, la calidad de los datos de ejemplo es uno muy importante. En teoría, tener más características debería resultar en una mayor potencia descriptiva o predictiva, sin embargo, la experiencia práctica ha demostrado que no siempre es éste el caso. Los problemas con una alta dimensionalidad, cantidad limitada de ejemplos disponibles y mucha información redundante o irrelevante son difíciles de tratar (3).

El caso de estudio claramente presentaba 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 tres años académicos.

4. Enfoque del Trabajo

Una vez que los datos han sido preprocesados utilizando los métodos descriptos en la sección anterior, la información es considerada una vista minable y está preparada para ser sometida a la técnica que permita establecer el modelo buscado.

La Minería de Datos presenta un amplio espectro de técnicas. La claridad de los resultados va a depender en gran medida de la técnica elegida, es por eso que este análisis previo resulta relevante.

La simple aplicación de una técnica de DM a una vista minable y el conocimiento previo del problema, no garantizan patrones expresivos, novedosos y útiles. Los algoritmos muchas veces ofrecen malos resultados debido a causas ajenas a su efectividad, ya sea porque no existe patrón en los datos o porque no se está usando la herramienta adecuada o porque el patrón es realmente difícil de encontrar.

Con esto en mente, se realizó una investigación bibliográfica de las tareas y métodos de DM con el objetivo de elegir un subconjunto de ellas en función de la utilidad que pudieran prestar al tratamiento de los datos con que se cuenta.

Existen dos tipos de tareas, las predictivas y las descriptivas. Se consideran predictivas a aquellas tareas que requieren de la obtención de un modelo capaz de dar una respuesta, en una etapa posterior, ante la presencia de información nueva. Según si la respuesta esperada es discreta o continua, se considera

que la tarea predictiva es una clasificación o una regresión, respectivamente. Un ejemplo de tarea de clasificación es obtener un modelo que a partir de la información de un nuevo alumno pueda clasificar su rendimiento como "Bueno", "Regular" o "Insuficiente". Un ejemplo de tarea de regresión es obtener un modelo que a partir de la información de un alumno nuevo pueda estimar la cantidad de años que demorará en recibirse.

Las tareas descriptivas buscan mostrar nuevas relaciones entre las variables y generalmente son utilizadas para mejorar el modelo. Su objetivo es describir los datos existentes. Entre las tareas descriptivas más frecuentes, puede mencionarse el agrupamiento (clustering), cuyo objetivo es obtener grupos o conjuntos entre los ejemplos, de manera que los elementos asignados al mismo grupo sean similares (9). A priori no se sabe ni cómo son los grupos ni cuantos hay, eso se determina con el proceso de aprendizaje. Una utilidad del agrupamiento reside en que utilizando la función obtenida con nuevos ejemplos se puede determinar a qué grupo pertenece el nuevo elemento y con eso indicar su comportamiento.

Otras tareas descriptivas son las correlaciones y factorizaciones, su objetivo es detectar si dos atributos numéricos están correlacionados linealmente o relacionados de algún otro modo. Su utilidad es la detección de atributos redundantes o dependientes y analizar la relevancia de atributos para hacer una selección entre ellos.

También encontramos dentro de las tareas descriptivas a las reglas de asociación que realizan un estudio similar al de correlaciones pero para atributos nominales.

En el caso de estudio la información de la que se dispone incluye algunos datos demográficos, económicos, sociales, familiares y académicos de los alumnos inscriptos en todas las carreras de grado de la UNRN desde su creación. Esta información podría ayudar a conocer los perfiles de los estudiantes que alberga la institución, en particular los perfiles de los alumnos desertores.

El mayor conocimiento del estudiante universitario es un tema que despierta el interés de las autoridades de la Universidad, ya que definir sus características aportaría a la organización la información mínima necesaria para implementar medidas paliativas de los fenómenos de abandono y desgranamiento. El agrupamiento de los datos permitió determinar que los factores sociales, económicos y familiares que describen a los estudiantes poseen una gran influencia en los índices de deserción. Por consiguiente, su análisis fue el primer paso en el camino hacia la disminución del riesgo de fracaso académico.

Estas consideraciones llevaron a orientar la investigación hacia las técnicas descriptivas, que permitirían resumir las características generales del conjunto de datos respecto a la información socio-económica y/o al rendimiento académico de los estudiantes de las diferentes cohortes en los años de vida de la UNRN.

5. Prueba de Concepto

Para esta primera aproximación al tratamiento del problema se seleccionaron las técnicas de agrupamiento, con la meta de caracterizar a los alumnos para ofrecer elementos de análisis.

Selección del subconjunto de datos.

Para la selección del primer subconjunto de datos a utilizar en las pruebas iniciales se tomaron en cuenta algunas consideraciones surgidas del conocimiento del dominio y del análisis realizado en la etapa de preparación de datos.

El conjunto de datos de que se disponía tenía a simple vista una composición heterogénea determinada por la existencia de instancias pertenecientes a alumnos ingresantes en el año 2012 de los cuales no se registra ninguna historia académica, lo que produce un vacío de información en varios atributos de la vista minable. Esta situación responde al hecho que la base de datos con la que se inicia este trabajo fue extraída desde la base del SIU-guaraní a principios de 2012, de manera que la ausencia de información es temporal. Como primera medida se exceptuaron esos datos, los registros ignorados en esta etapa del

análisis podrán ser completados en años subsiguientes con los valores académicos generados y probablemente utilizados como fuente de datos para control y validación de resultados y para el desarrollo de nuevos modelos.

A partir de los datos resultantes, se decidió agrupar 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.

La implementación de todos los métodos utilizados a lo largo del trabajo se resolvió mediante la utilización de RapidMiner 5.2 (7), herramienta *open source* de minería de datos.

Como conjunto de datos de entrada se usó la vista minable generada en la sección de preparación de datos, previa selección de los registros con *estado = 'Abandono'*.

Se realizaron las tareas de transformación de datos necesarias para la aplicación del algoritmo, como la numerización las variables de entrada y se aplicó una normalización de rango sobre todos los atributos utilizando el método de transformación z, mediante el operador *Normalize* de RapidMiner.

Se realizaron varios intentos de aplicación del algoritmo de agrupamiento k-medias (operador *k-means* de RapidMiner) al conjunto de datos con valores diferentes de *k*, obteniendo finalmente 5 grupos.

Luego se enfrentó la tarea de describir los grupos obtenidos a través de sus centroides y calcular los valores frecuentes en los grupos (5). En todos los casos se utilizó el conocimiento del dominio para guiar la descripción, pero los resultados no fueron satisfactorios dada la cantidad de atributos intervinientes.

Las pruebas realizadas aplicando k-medias pusieron en evidencia la gran dimensionalidad del problema, que oscurece la interpretación de los agrupamientos obtenidos. Dada la cantidad de atributos involucrados (todos los de la vista minable inicial), no es posible encontrar un conjunto de clusters descriptivo de los datos de entrada.

La selección de atributos puede ser guiada por el conocimiento del dominio o por técnicas específicas de DM. En este punto surgió la necesidad de utilizar las herramientas de la minería de datos para guiar la selección de un subconjunto de características (atributos) que sean relevantes para el problema.

Por esta razón se dejó en suspenso la aplicación de los métodos de agrupamiento y se enfocó la tarea en la utilización de técnicas que permitieran visualizar el conjunto de atributos adecuado para la aplicación de dichos métodos.

Selección de características relevantes.

El objetivo en este punto era encontrar un subconjunto de atributos del conjunto total inicial, que incluyera aquellos relevantes para la tarea de agrupamiento. Este nuevo marco de trabajo se enfrentó con un problema: las técnicas conocidas de selección de características parten de un conjunto de datos etiquetados, es decir, un grupo de registros en el cual cada uno de ellos lleva una indicación de la clase a la cual pertenece. Luego los procesos selectivos apuntan a obtener las características más correlacionadas con la clase. No es posible calcular correlación o relevancia si no existe un atributo clase en los datos.

La solución a este problema se halló en el resultado del agrupamiento inicial obtenido con k-medias. Si bien no fue posible describir los grupos obtenidos de manera aceptable, sí se pudo obtener un atributo de clase (dado por el grupo al cual cada ejemplo pertenece), de manera que la entrada a los algoritmos de selección de características es el resultado del paso anterior, el agrupamiento en cinco grupos de los alumnos desertores.

Una vez determinado el conjunto de datos de entrada, se procedió a la transformación del espacio de características mediante dos esquemas. El primero de ellos hace uso de un proceso de selección del tipo S*election Forward* que toma en cuenta la performance de un determinado modelo de aprendizaje para realizar la validación de los conjuntos de características. El uso de un algoritmo inductivo posiciona al método dentro de los

procesos *wrapper* (4). Como algoritmo de validación se utilizó un agrupamiento del tipo k-medias, con k = 5. Para la implementación del procedimiento se utilizó el operador *Optimize Selection* de RapidMiner, con la parametrización adecuada (8). El segundo método de selección implementado está enfocado a la selección genética de características (a través de mutación y cruce) (1), que no solo intenta maximizar la performance del conjunto de características sino también minimizar el número de ellas. Dado que no utiliza un algoritmo inductivo determinado para la evaluación, también es conocido como selección multiobjetivo, es de propósito general y se adapta a los casos de poco conocimiento del dominio del problema. Para la evaluación utiliza el método CFS (*Correlation-based Features Selection*) (2). La implementación se realizó con los operadores *Optimize Selection (Evolutionary)* y *Performance (CFS)* de RapidMiner (8).

Ambos métodos proporcionaron subconjuntos de características bastante similares entre sí (ver tablas 1 y 2.)

Validación de características seleccionadas. Árbol de decisión.

La etapa de selección de características es de suma importancia para la obtención de buenos resultados, de forma tal que antes de utilizar el subconjunto de datos obtenido en el procesamiento anterior, se recurrió a una instancia más de validación para el subconjunto de características Existen algoritmos de aprendizaje automático que están diseñados para aprender cuales son los atributos más apropiados para tomar decisiones. Por ejemplo, los árboles de decisión eligen el atributo más prometedor para llevar a cabo la división en cada nodo interno, y no deberían seleccionar atributos irrelevantes o carentes de utilidad. A medida que se avanza en la técnica "divide y vencerás" propia de estos métodos (a medida que se va descendiendo en los niveles del árbol), la cantidad de ejemplos involucrados disminuye y la posibilidad de seleccionar atributos irrelevantes para la división aumenta. Los niveles superiores del árbol, entonces, deben representar el conjunto de atributos de mayor interés para el problema.

estado_civil = soltero
padre_vive = S
alu_tec_int = tiene Internet en la casa
rel_trab_carrera
alu_trab_remmon
sede = Valle Medio y Rio Colorado
lugar_nacimiento = NEUQUEN
colegio_secundario = N
titulo_secundario = PER.MERC:
sit_laboral_padre = DesOcupado
sit_laboral_madre = No Informado
alu_trab_sitimp = Relación de dependencia
alu_trab_sitimp = no trabaja
alu_trab_sitimp = Monotributista
alu_trab_futtip = No trabajaré
alu_trab_futtip = Desconoce
alu_trab_futtip = Cuenta Propia
anio_nacim
cant_fami_cargo
cant_hijos_alum

Tabla 1. Lista de atributos seleccionados por método wrapper.

En base a esta idea, como medida de validación de los atributos seleccionados, se construyó un árbol de decisión. Para esta tarea se eligió el algoritmo C4.5 (6) implementado por el operador *W-J48* de la extensión Weka de RapidMiner (ver figura 1).

Se puede apreciar que en los niveles superiores del árbol se encuentran los atributos *sit_laboral_madre, alu_trab_sitimp, alu_trab_remmon, cant_hijos_alum, anio_nacim, estado_civil, cant_fami_cargo, sit_laboral_padre, rel_trab_carrera*, todos atributos presentes en los subconjuntos generados por los algoritmos de selección. Luego de estas comprobaciones y en correspondencia con la decisión ya tomada de investigar el problema a través de métodos de agrupamiento, se optó por el subconjunto de características obtenido con la metodología *Selection Forward*, avalada por el resto de los procesos selectivos ejecutados.

estado_civil = Soltero
padre_vive = S
alu_beca = necesita beca
rel_trab_carrera
alu_trab_remmon
alu_trab_futhor
sede = Rectorado
lugar_nacimiento = ADOLFO ALSINA
titulo_secundario = BACHILLER
sit_laboral_padre = DesOcupado
sit_laboral_madre = No informado
alu_trab_sitimp = Relación de dependencia
alu_trab_sitimp = no trabaja
alu_trab_sitimp = Monotributista
alu_trab_futtip = Obrero o empleado (asalariado)
alu_trab_futtip = Cuenta Propia
anio_egreso_sec
cant_fami_cargo
cant_hijos_alum

Tabla 2. Lista de atributos seleccionados por método genético.

Aplicación del modelo para las características seleccionadas.

El siguiente paso fue la ejecución de k-medias para los atributos seleccionados, nuevamente con k = 5. Los datos de entrada fueron esta vez los registros pertenecientes a alumnos desertores, con las mismas transformaciones descriptas en la primera corrida del método, pero solo aplicadas a los atributos del subconjunto seleccionado. El universo de ejemplos fue mismo que la primer aplicación: los alumnos con *estado = 'Abandono'*, la diferencia radica en los atributos que describen a cada ejemplo.

El proceso de reducción de dimensionalidad determinó que los datos relevantes para agrupar a los alumnos desertores son variables de tipo socio-económicas, como la edad, el estado civil, las cargas familiares, la situación laboral actual y futura del alumno y la de sus padres. Un dato sobresaliente es la ausencia de atributos académicos en el grupo de relevancia.

Una vez aplicado el método, el resultado de la asignación a grupos de esta ejecución se comparó con el resultado de la ejecución anterior, determinando que menos de un 10% de los ejemplos se movieron de grupo, lo que indicó que el criterio de agrupamiento se conservó a pesar de la reducción de características.

```
Sit_laboral_madre = No informado <= -0.2
|   Alu_trab_sitimp = Monotributista <= -0.3
|   |   Alu_trab_remmon <= -1.1
|   |   |   Cant_hijos_alum <= 0.2: CLUSTER2
|   |   |   Cant_hijos_alum > 0.2
|   |   |   |   anio_nacim <= -0.3
|   |   |   |   |   estado_civil = Soltero <= -1.5
|   |   |   |   |   |   Padres_prop_viv = Propia <= -1.5: CLUSTER0
|   |   |   |   |   |   Padres_prop_viv = Propia > -1.5: CLUSTER2
|   |   |   |   |   estado_civil = Soltero > -1.5: CLUSTER2
|   |   |   |   anio_nacim > -0.3: CLUSTER2
|   |   Alu_trab_remmon > -1.1
|   |   |   Cant_hijos_alum <= 0.2
|   |   |   |   estado_civil = Soltero <= -1.5
|   |   |   |   |   Cant_hijos_alum <= -0.7
|   |   |   |   |   |   anio_nacim <= -0.6: CLUSTER0
|   |   |   |   |   |   anio_nacim > -0.6: CLUSTER1
|   |   |   |   |   Cant_hijos_alum > -0.7
|   |   |   |   |   |   Sit_laboral_padre = Ocupado <= 1:CLUSTER0
|   |   |   |   |   |   Sit_laboral_padre = Ocupado > -1: CLUSTER1
|   |   |   |   estado_civil = Soltero > -1.5
|   |   |   |   |   anio_nacim <= -0.6
|   |   |   |   |   |   Cant_fami_cargo <= -0.7: CLUSTER1
|   |   |   |   |   |   Cant_fami_cargo > -0.7
|   |   |   |   |   |   |   anio_nacim <= -1.1: CLUSTER0
|   |   |   |   |   |   |   anio_nacim > -1.1: CLUSTER1
|   |   |   |   |   anio_nacim > -0.6: CLUSTER1
|   |   |   Cant_hijos_alum > 0.2
|   |   |   |   estado_civil = Soltero <= -1.5: CLUSTER0
|   |   |   |   estado_civil = Soltero > -1.5
|   |   |   |   |   Sit_laboral_padre = Ocupado <= -1: CLUSTER0
|   |   |   |   |   Sit_laboral_padre = Ocupado > -1
|   |   |   |   |   |   Rel_trab_carrera <= 0.4: CLUSTER1
|   |   |   |   |   |   Rel_trab_carrera > 0.4: CLUSTER0
|   Alu_trab_sitimp = Monotributista > -0.3: CLUSTER4
Sit_laboral_madre = No informado > -0.2: CLUSTER3
Number of Leaves :          19
Size of the tree :  37
```

Figura 1. Arbol de decisión.

Los 5 grupos resultantes debían ahora ser descriptos y representados de manera que aportaran valor al tratamiento del problema. Esta es la tarea que se vio obstaculizada por la alta dimensionalidad de los ejemplos, se esperaba que en esta instancia la descripción de los grupos se clarificara al estar determinada por un número notablemente menor de variables.

La figura 2 muestra la representación gráfica provista por RapidMiner de los centroides resultantes.

Descripción de perfiles obtenidos.

Se utilizaron los centroides y los valores frecuentes en cada grupo para representarlo, considerando las frecuencias de valores a +/-

una desviación estándar del valor del centroide para los atributos. Una vez realizado este trabajo se le asignó un nombre descriptivo a cada grupo:

Mayores Relación de Dependencia (Cluster 0): 848 alumnos. Tienen una edad promedio de 44 años, trabajan en relación de dependencia. Su trabajo está relacionado parcial o totalmente con la carrera que cursan, poseen cargas familiares, no son solteros. Ganan más de 2000 $.

Mayores Monotributistas (Cluster 4): 354 alumnos. La edad promedio de este grupo es de 40 años, son en su mayoría solteros, sus padres viven, trabajan como monotributistas y más de la mitad tiene cargas familiares.

Mayores sin Información (Cluster 3): 105 alumnos. Este grupo de relativamente pequeña cardinalidad se caracteriza por tener muchos atributos no informados, lo que solo permite afirmar que son de edad promedio 40 años y no tienen padres.

Menores Trabajan (Cluster 1): 1259 alumnos. El promedio de edad del grupo es de 30 años, trabajan en relación de dependencia con un sueldo menor a 2000$ en tareas no relacionadas o solo relacionadas parcialmente con sus carreras, no tienen cargas familiares, son solteros, sus padres trabajan.

Menores no Trabajan (Cluster 2): 1163 alumnos. Son los más jóvenes, con promedio de 26 años, no trabajan, sus padres viven, no tienen cargas familiares y son solteros.

La selección de características permitió encontrar descripciones concretas de los grupos que caracterizan a los alumnos que contienen y los diferencian claramente de los pertenecientes a otros grupos.

Agrupamiento para la obtención de perfiles del alumno no desertor.

La mera segmentación en grupos de los alumnos que abandonan permitía tener un mayor conocimiento de los subgrupos que componen la clase de interés (alumnos desertores), pero no permitía compararla con la clase de alumnos que continúan con sus carreras. Una opción que podía colaborar a la integración de conceptos era la realización de un agrupamiento sobre los alumnos que continúan cursando, utilizando una vista minable con los mismos atributos seleccionados.

Con la idea de evaluar la posibilidad de una comparación entre los alumnos desertores y los que no lo son (al menos hasta el momento), se prepararon los datos de estos alumnos con los mismos criterios de numerización y normalización utilizados anteriormente y se ejecutó sobre ellos un método de agrupamiento k-medias con k = 5. La representación gráfica de los centroides obtenidos se muestra en la figura 3.

Al describir los grupos obtenidos, claramente se encontraron grupos “comparables” con los anteriores, los que se caracterizan a continuación:

Mayores Relación de Dependencia (Cluster 3): 511 alumnos. Tienen una edad promedio de 44 años, trabajan en relación de dependencia. Su trabajo está relacionado parcial o totalmente con la carrera que cursan, poseen cargas familiares, no son solteros. Ganan más de 2000 $.

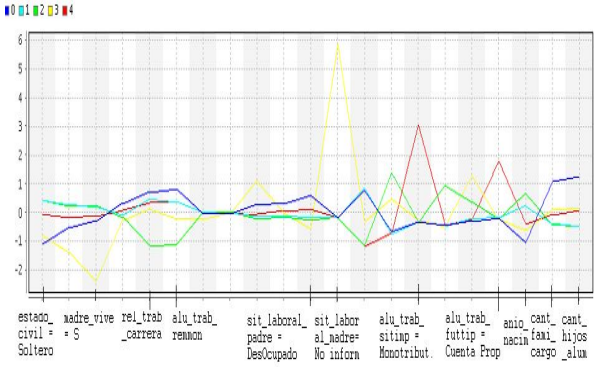

Figura 2. Centroides clusters Abandonos.

Mayores Monotributistas (Cluster 4): 220 alumnos. La edad promedio de este grupo es de 37 años, son en su mayoría solteros, sus padres viven, trabajan como monotributistas y algo menos de la mitad tiene cargas familiares.

Mayores sin Información (Cluster 2): 85 alumnos. Este grupo de relativamente pequeña cardinalidad se caracteriza por tener muchos atributos no informados, lo que solo permite afirmar que son de edad promedio 37 años y no tienen padres.

Menores Trabajan (Cluster 0): 966 alumnos. El promedio de edad del grupo es de 29 años, trabajan en relación de dependencia con un sueldo menor a 2000$ en tareas no relacionadas o solo relacionadas parcialmente con sus carreras, no tienen cargas familiares, son solteros, sus padres trabajan.
Menores no Trabajan (Cluster 1): 1870 alumnos. Son los más jóvenes, con promedio de 24 años, no trabajan, sus padres viven, no tienen cargas familiares y son solteros.

6. Interpretación de Resultados

La descripción y caracterización de los grupos emergentes del trabajo previo permite analizar el fenómeno de deserción de las cohortes de la UNRN, desde su creación hasta el año 2011 inclusive, desde un punto de vista innovador.
Se han podido establecer grupos que describen a los alumnos que han abandonado y a los que no lo han hecho, y se ha expuesto la correspondencia entre los grupos análogos dentro de cada clase (donde la clase, en esta instancia evaluativa, es binaria e indica el estado o no de abandono).
Realizada la correspondencia anterior, es posible avanzar en la interpretación de estos grupos de manera conjunta, buscando que puedan aportar información útil para el tratamiento de la deserción.
Un simple cálculo porcentual para cada grupo indica claramente en cuales clusters hay mayor incidencia desertora. De esta forma se obtienen los siguientes porcentajes:
Mayores Relación de dependencia. Abandona el ***62%***
Mayores Monotributistas. Abandona el ***62%***
Mayores sin Información. Abandona el **55%**
Menores Trabajan. Abandona el ***56%***
Menores no Trabajan. Abandona el ***38%***
Es claro desde aquí que los grupos de menor edad, solteros sin cargas familiares y que no trabajan, tienen el menor índice de abandono, mientras que los mayores índices de abandono se perciben entre los grupos de mayor edad, con más cargas familiares y que trabajan. Se puede ver también que la variable del trabajo tiene una alta incidencia en el abandono, incluso en grupos más jóvenes.

Ahora bien, establecida la incidencia laboral en la pérdida de continuidad en el estudio, queda un grupo muy numeroso de jóvenes que no trabajan y que de todas maneras abandonan sus carreras (1163 alumnos). Es quizá el grupo sobre el cual resta más trabajo por hacer.

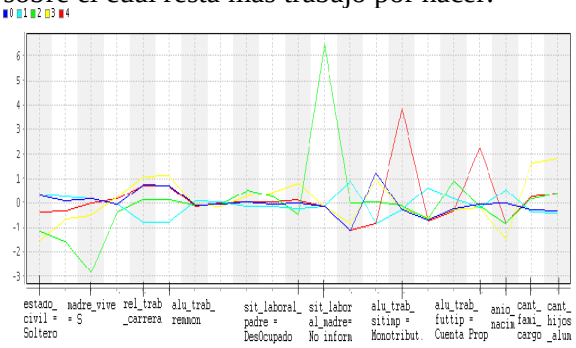

Figura 3. Centroides clusters Cursan.

Probablemente sea de gran utilidad determinar las posibles causas de deserción en esta franja de estudiantes. Más aún, es de interés poder determinar si las causas están vinculadas a factores intra-institucionales en lugar de vincularse a factores externos como las cargas familiares y el trabajo. El descubrimiento de estos factores, aún ocultos, podría determinar los cambios en la Institución Universitaria que favorezcan la disminución del número de abandonos.

Trabajos futuros.
Es notorio, desde la interpretación de resultados, que los logros alcanzados son solo la etapa preliminar de los estudios que pueden realizarse utilizando técnicas de DM. La continuidad de la investigación pretende avanzar en la utilización de otras técnicas sobre el mismo conjunto de datos y sobre datos actualizados de la misma Universidad. Se espera poder aplicar técnicas de aprendizaje supervisado para obtener modelos que puedan ayudar a predecir el fenómeno de deserción universitaria y definir estrategias de intervención.
Los próximos objetivos probablemente incluyan la obtención de nuevos modelos, utilizando técnicas de DM aún no abordadas en este desarrollo, que permitan avanzar en la determinación de las variables incidentes en el abandono de los alumnos, proporcionando

herramientas para la implementación de decisiones que disminuyan el riesgo de deserción.

7. Conclusiones

Se inició esta investigación con el objetivo de estudiar las técnicas de DM y su aplicabilidad al análisis de la deserción de alumnos universitarios de la UNRN.

Luego del camino recorrido se pueden destacar los logros siguientes:

- Se conoció y analizó el conjunto de dimensiones que forman parte del dominio del problema y los aspectos que interactúan para caracterizar la población objeto de estudio.
- Se prepararon los datos disponibles para aplicar algunos modelos de minería de datos.
- Se realizó una investigación bibliográfica sobre la metodología del proceso de extracción de conocimiento en grandes bases de datos, seleccionando algunas técnicas y algoritmos para abordar el problema.
- Se recorrió el proceso metodológico sugerido por la bibliografía, solucionando los problemas encontrados, dejando en claro la naturaleza iterativa del mismo.
- Se corroboró la factibilidad del uso de la tecnología de DM en la extracción de conocimiento para el caso de estudio.
- Se llevó a término una prueba de concepto que arroja información preliminar relevante respecto a la problemática del abandono.
- Se consiguió describir los perfiles de los estudiantes aportando información útil en relación a su composición socio-económica y su permanencia en el ámbito universitario, demostrando también que las técnicas elegidas se adaptan al objetivo planteado.

Esto es apenas la punta del iceberg. Se estuvo arañando la superficie de la mina y se encontraron temas a investigar, lo cual indica que perseverando en este camino de Minería de Datos se pueden encontrar resultados que favorezcan el diseño de modelos útiles para el abordaje del problema.

Bibliografía

1. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st ed.
2. Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand.
3. Hernández Orallo, J., Ramírez Quintana, M., and Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. Ed. Pearson.
4. Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artif. Intell., 97(1-2):273–324.
5. Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Data-Centric Systems and Applications. Springer.
6. Quinlan, R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
7. RapidMiner (2012). Rapid miner. http://http://rapid-i.com/content/view/181/190. [Ultimo acceso : 18-Nov-2012].
8. Tito, L. and Mullicundo, F. (2010). Rapidminer. tutorial on-line + operadores. http://es.scribd.com/doc/78886734/Rapid-Miner-Tutorial-Online-Ope-Rad-Ores.
9. 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.