Exploración de Datos Académicos a través de la aplicación de Técnicas de Minería de Datos en Weka

Autora: Eckert Karina Beatriz

Director: Ing. Suénaga Roberto

Universidad Gastón Dachary – Misiones Argentina
karinaeck@gmail.com - rsuenaga@ugd.edu.ar

Ingeniería en Informática con Orientación en Sistemas de Información
Año 2012

adfa, p. 1, 2011.

Exploración de Datos Académicos a través de la Aplicación de Técnicas de Minería de Datos en Weka

Abstract. La presente investigación expone la aplicación del proceso denominado Descubrimiento de Conocimiento en Bases de Datos (KDD), conocido como Minería de Datos (MD), sobre la información académica de la Universidad Gastón Dachary (UGD). Dicho proceso consiste en una serie de etapas iterativas que incluyen el pre y post procesamiento de datos, hasta obtener conocimiento nuevo. Para ello, se realizaron numerosas selecciones y depuraciones de datos, utilización de diferentes criterios de representación y aplicación de diferentes técnicas y algoritmos.

La fuente de datos proviene de la información proporcionada al ingreso (personales y antecedentes educativos) y durante el lapso de sus estudios; con la debida protección de datos personales, creando una vista minable con las características de las titulaciones seleccionadas, una colección de individuos sobre los cuales se realizó el estudio para extraer conocimiento útil en lo que se refiere a rendimiento académico, correspondiente un periodo de 10 años (1999-2009).

Se ha detectado mediante ciertos algoritmos correspondientes a las técnicas de asociación, clustering, selección de atributos y clasificación, que existen tendencias y relaciones entre los datos pertenecientes a los departamentos evaluados; existiendo particularidades y coincidencias entre éstos.

Keywords: Knowledge Discovery in Databases (KDD), Educational Data Mining (EDM), WEKA

Introducción.

Actualmente la sociedad se encuentra en la denominada era de la información, donde los sistemas informáticos tienen cada vez mayor capacidad, y este incremento de almacenamiento tiene un efecto realmente interesante, dado por el bajo costo que tiene guardar los datos del funcionamiento de los procesos, o sub sistemas de una organización en las bases de datos (en el sentido más amplio del término), lo que conlleva a que éstas crezcan hasta límites inesperados. Empero, contrariamente a lo se espera, esta expansión de datos no supone un aumento de conocimiento, puesto que procesarlos con los métodos clásicos resulta ser en muchos casos imposible o sumamente tedioso y con resultados superficiales e insatisfactorios.

En los últimos años han surgido una serie de herramientas de MD, tales como Weka, RapidMiner, SPSS Clementine, entre otras, basadas en técnicas que facilitan el procesamiento avanzado de datos y permiten realizar un análisis en profundidad de los mismos de manera automática [1] [2].

El proceso completo de extracción del conocimiento oculto en los datos almacenados en las base de datos se denomina KDD (*Knowledge Discovery in Databases*) y más en concreto mediante las técnicas utilizadas en una de sus fases, denominada

adfa, p. 1, 2011.

Minería de Datos (MD), es la encargada de la identificar este conocimiento oculto. El Proceso de KDD requiere una serie de condiciones para su correcta utilización, pero resulta sumamente útil su aplicación en los sistemas tradicionales.

Las instituciones educativas no escapan a este tipo de problemáticas, la limitación del conocimiento de los datos, puede generar incongruencias en la toma de decisiones, en las proyecciones y en las nuevas políticas que se requieren para mejorar el ámbito educativo. Es evidente la necesidad de disponer de un sistema de gestión que permita tomar decisiones y elaborar estrategias a partir del conocimiento oportuno, ya que esto no solo incide directamente sobre la funcionalidad de los departamentos académicos, u otras cuestiones internas, sino que también podrían incidir sobre actividades como las evaluaciones y acreditaciones de instituciones y carreras. Por lo tanto es indispensable considerar la aplicación de un modelo de gestión del conocimiento en estructuras organizacionales de este tipo.

Esta investigación desde el punto de vista metodológico, es tanto descriptiva como experimental. Se describen las distintas técnicas y algoritmos de procesamiento de información mediante herramientas de MD, así como también se describen los procesos que se deben llevar a cabo para depurar, clasificar, procesar y obtener información significativa de una base de datos educativa. Además es un trabajo experimental debido a que se utilizan datos reales, los cuales son acondicionados y procesados mediante algoritmos adecuados (árboles de decisión, reglas, etc.) a las técnicas de MD (preprocesamiento, clasificación, segmentación, asociación, selección de atributos y visualización) y en base a los resultados obtenidos se reitera el proceso, que en sí es de carácter iterativo e interactivo, para refinar el conocimiento obtenido. Además de detectar patrones y relaciones entre los datos, que sirva de referencia para la toma de decisión en el ámbito de la gestión académica; se buscó desarrollar una experiencia tanto en el proceso de KDD, MD en el ámbito de la educación, así como también en la utilización de herramientas de MD, técnicas y algoritmos de procesamiento de información.

El proceso de KDD es una disciplina emergente, con una serie de ventajas, por un lado, es punto de encuentro entre los investigadores e integrantes de las organizaciones, y por otro, permite dar un mayor prestigio a las instituciones, lo cual, en el ámbito educativo está relacionado a la calidad académica.

Objetivos.

1 General.

Determinar patrones y relaciones entre datos de la trayectoria académica de los alumnos de la UGD utilizando técnicas de Minería de Datos (MD).

Se utilizarán técnicas de MD en el ámbito de la Universidad Gastón Dachary para conocer las variables que impactan el desempeño académico de los estudiantes. De esta manera se espera estimar cuales son las variables que tendrán mayor impacto en el desempeño, la permanencia y la graduación de los estudiantes en la institución.

Los patrones corresponden a comportamientos comunes en cuanto a dificultad/facilidad para aprobar materias, etapas o tramos de una carrera; mientras que las relaciones se refieren a la vinculación o asociación entre datos, como por ejemplo: alumnos procedentes de determinados establecimientos secundarios poseen dificultades en determinadas áreas del conocimiento (el interés estará dado en relaciones entre datos múltiples).

2 Específicos.

- Describir el proceso KDD, técnicas, procesos y herramientas de MD que se utilizan en el ámbito educativo.
- Seleccionar y acondicionar los datos datos relacionados con el rendimiento académico de los estudiantes de acuerdo a los requerimientos de la herramienta de MD.
- Utilizar WEKA como herramienta de MD para detectar patrones y relaciones entre los datos de la trayectoria académica de los estudiantes.
- Identificar las variables que impactan el desempeño académico.
- Explicar los resultados obtenidos y elaborar recomendaciones sobre los posibles usos de los resultados obtenidos y las características de las fuentes de información.

Desarrollo Experimental.

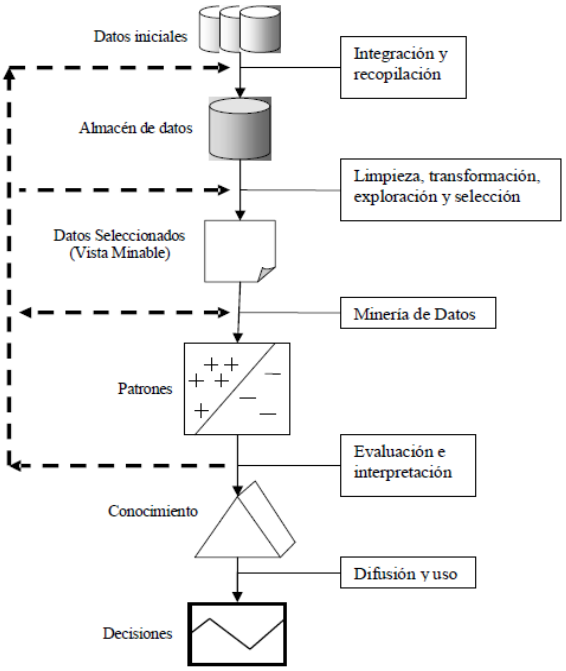

Fig. 1. Fases del Proceso de KDD

En lo que se refiere específicamente al desarrollo experimental de la investigación, se siguió la estructura del Proceso KDD, que costa de cinco fases como se indica en la Figura 1 [3].

Durante el desarrollo del proceso de KDD, suele ser necesario interrumpir en algún punto de las fases del proceso y volver a comenzar en alguno de los pasos anteriores, siendo así un proceso iterativo e interactivo necesario para lograr una alta calidad del conocimiento a descubrir [4].

1 Integración y Recopilación de Datos.

1.1 Análisis de la Base de Datos de la Universidad.

Partiendo de los datos operacionales provenientes de la Base de Datos de la UGD, se llevó a cabo una descripción del dominio de la aplicación que se refiere a la información registrada de índole académico de los estudiantes de dicha institución. Toda la información con la que cuenta la UGD, esta almacenada en alrededor de 50 tablas de una base de datos relacional.

1.2 Consolidación de los Datos.

Durante el análisis de los datos se encontraron los siguientes problemas:

- Faltante de registros en tablas.
- Registros con datos cargados incorrectamente.
- Carencia de explicaciones referente al contenido de las tablas existentes.

Se debió realizar un control y limpieza exhaustiva de los datos para hallar una coherencia completa de las tablas y obviar datos irrelevantes para el trabajo.

2 Filtrado de Datos: limpieza, transformación, exploración y selección de datos.

2.1 Limpieza de Datos.

Se realizó un análisis de la Base de Datos en busca de valores erróneos, faltantes, no uniformes o irrelevantes, donde se encontraron casos como los siguientes:

- Datos Faltantes: atributos no obligatorios pero con información posiblemente importante para investigación (vacíos o nulos).
- Datos Erróneos: atributos con valores por defecto que pueden influir en el resultado de la investigación.
- Datos No Uniformes: el mismo valor del atributo se representa de varias maneras distintas. Para solucionar estos problemas se definieron criterios para unificar valores de contenidos y se procedió a la corrección.

2.2 Transformación de Datos.

Actividades como la creación de atributos derivados, discretizaciones, transformación de los valores existentes o generación de nuevos (sintácticas o que tengan que ver con el orden en que son presentados los atributos).

2.3 Análisis Explorativo.

Útil para identificar patrones, distribuciones de valores en un conjunto de datos, y así observar el comportamiento y la calidad del conjunto de datos previamente a la aplicación de técnicas y algoritmos de MD.

Fig. 2. Atributos y distribución gráfica de sus valores

En la Figura 2 se puede apreciar, de manera general, el comportamiento de los atributos y su distribución gráfica de valores gracias a la herramienta de MD Weka (Visualize All del Preprocess del Explorer). Al realizar un análisis de las distribuciones de valores, se ha decidido modificar y descartar algunos de ellos.

2.4 Selección de Datos a Minar.

Tras varias iteraciones del proceso de KDD y en base a los resultados obtenidos en las pruebas previas, se decide dividir la muestra total en los dos departamentos académicos principales seleccionados para la investigación, tomando al departamento de Informática por un lado y al de Administración por el otro. La razón inicial de dicha división, fue para resolver problemas y dificultades en la comprensión de los resultados, por tratarse de estructuras de carreras diferentes.

La UGD posee una estructura departamental, es decir que las carreras se desarrollan agrupadas por disciplinas bajo la estructura de departamentos, equivalentes a facultades en otras universidades.

3 Minería de Datos, Interpretación, Evaluación y Extracción de Conocimiento.

En este punto, tenemos ambas vistas minables dispuestas y adecuadas para aplicar las técnicas y algoritmos de MD sobre ellas, obteniendo modelos de referencia que puedan servir para la toma de decisiones. Existe la necesidad de determinar en qué nivel de madurez se encuentran dichos modelos y si son adecuados para resolver el o los problemas planteados, por lo que son revisados, interpretados y evaluados, para extraer el conocimiento significativo.

Las técnicas de MD, que se dividen en dos categorías principales según el objetivo del modelo, descriptivas y predictivas.

3.1 Técnicas Descriptivas de Minería de Datos.

Asociación.

El algoritmo más popular y utilizado es *Apriori*; el cual genera conjuntos de reglas, ordenadas de acuerdo a la medida de confianza entre paréntesis después de cada regla. El número que le sigue al antecedente muestra cuantos casos cumplen el antecedente, y el número después del consecuente indica cuantas instancias satisfacen toda la regla (es una regla de "apoyo") [2].

Muchas veces los resultados obtenidos deben tomarse como tendencias, sobre todo en los casos donde la precisión (nivel de confianza) y el número de instancias son bajos con respecto a la muestra (cobertura/soporte). Siguiendo con las etapas del proceso de KDD, al analizar, interpretar y evaluar los resultados obtenidos en base a las reglas de asociación más significativas, se destacan los siguientes aspectos:

Promedios.

Los promedios de exámenes finales aprobados del 1° y 2° año inciden directamente sobre los obtenidos durante la totalidad de carrera. Para ambos departamentos, los promedios de materias aprobadas superiores a 6 mantienen dicha tendencia; adicionalmente para el Dpto. de Administración se detectó la relación existe entre los promedios generales y el éxito de llegar a los últimos años de la carrera para valores entre 5 y 6. Algunas de las reglas obtenidas se exponen en la Tabla 1.

Table 1. Reglas de Asociación com Algoritmo Apriori para los Promedios

Reglas de Asociación	Precisión
10. PromAp1=(6-7] 416 ==> PromAproC=(6-7] 342	conf: 0.83 (Inf)
16. PromAp2=(6-7] 261 ==> PromAproC=(6-7] 204	conf: 0.78 (Inf)
13. PromAp1=(6-7] 830 ==> PromAproC=(6-7] 628	conf: 0.76 (Adm)
20. PromAp1=(7-10] 301 ==> PromAproC=(7-10] 229	conf: 0.76 (Inf)
5. PromAp1=(7-10] 648 ==> PromAproC=(7-10] 558	conf: 0.86 (Adm)

Condiciones.

La mayoría de las reglas obtenidas refieren a la condición de baja definitiva, relacionado con el bajo número de asignaturas aprobadas y cursadas, o altos grados de fracasos en cursados (libres o re-cursantes). Para la baja temporal (pago de la cuota mensual suspendido, que puede corresponder con los alumnos en instancia de tesis) se identificó que ocurren en los últimos años de las carreras, especialmente para el Dpto. de Informática (75% de los casos). Varias reglas pueden observarse en la Tabla 2.

Table 2. Reglas de Asociación con Algoritmo Apriori para las Condiciones

Reglas de Asociación	Precisión
21. Condicion=B_DEF 454 ==> AñoAcad=1°a3° 343	conf:0.76 (Inf)
5. AñoAcad=1°a3° Loc=OTRAS 254 ==>Condicion=B_DEF 228	conf:0.90 (Inf)
9. AñoAcad=1°a3° 1°Apr=[1-3] 234 ==>Condicion=B_DEF 196	conf:0.84 (Inf)
1. AñoAcad=1°a3° Loc=OTRAS 473==>Condicion=B_DEF 440	conf:0.93(Adm)
11. AñoAcad=1°a3° Apro2=[0-1] 541 ==>Condicion=B_DEF 42	conf:0.79(Adm)
16.Apro2=[0-1] Condicion=B_DEF 583==>AñoAcad=1°a3° 425	conf:0.73(Adm)
17.AñoAcad=1°a3°Condicion=B_DEF 616==>Loc=OTRAS 440	conf:0.71(Adm)

Años Académicos.

Existe una relación directamente proporcional entre el número de materias aprobadas y cursadas, y el grado de éxito o fracaso para llegar a los últimos años de las carreras. Seguido en la Tabla 3, algunas de las reglas obtenidas por el algoritmo Apriori.

Table 3. Reglas de Asociación con Algoritmo Apriori para los Años Académicos

Reglas de Asociación	Precisión
1. 1°Apr=[1-3] Apro2=[0-1] 205 ==> AñoAcad=1°a3° 192	conf: 0.94 (Inf)
7. 1°Apr=[1-3] Condicion=B_DEF 225 ==> AñoAcad=1°a3° 196	conf: 0.87 (Inf)
13. 1°Apr=[1-3] 293 ==> AñoAcad=1°a3° 234	conf: 0.80 (Inf)
14. 1°Apr=[8-9] 272 ==> AñoAcad=4°o5° 217	conf: 0.80 (Inf)
3. 2°Apr=[4-7] 530 ==> AñoAcad=4°o5° 471	conf: 0.89(Adm)
2. Apro2=[6-inf) 511 ==> AñoAcad=4°o5° 463	conf: 0.91(Adm)
4. Curs2=[10-inf) 493 ==> AñoAcad=4°o5° 431	conf: 0.87(Adm)
6. FracC1=[0] 500 ==> AñoAcad=4°o5° 422	conf: 0.84(Adm)
19. EdadIng=[23o+] 828 ==> AñoAcad=4°o5° 580	conf: 0.70(Adm)

Segmentación o Agrupamiento.

Agrupamiento Numérico: Algoritmo EM.

EM proviene de la estadística y es más elaborado que otros, con el coste de que requiere mayor cantidad de operaciones, lo que lo hace lento, aún así tiene la ventaja adicional de buscar automáticamente el número de segmentos apropiados a la muestra [5]. Se han generado 21 clusters, como se demuestra en la Tabla 4. Las distribuciones de instancias respecto a los clusters son bastante homogéneas, que van desde el 1% al 8% para el Dpto. de Informática y de 2% al 7% para el de Administración.

Table 4. Segmentación Algoritmo EM (Clustered Instances)

Dpto. Informática			Dpto. Administración		
29 (3%)	73 (8%)	35 (4%)	62 (3%)	117 6%)	62 (3%)
75 (8%)	20 (2%)	44 (5%)	93 (5%)	83 (4%)	91 (4%)
28 (3%)	61 (6%)	52 (5%)	83 (4%)	147(7%)	78 (4%)
57 (6%)	44 (5%)	13 (1%)	85 (4%)	38 (2%)	107(5%)
51 (5%)	52 (5%)	49 (5%)	72 (4%)	63 (3%)	115 (6%)
36 (4%)	35 (4%)	65 (7%)	149 (7%)	140 (7%)	80 (4%)
48 (5%)	61 (6%)	30 (3%)	148(7%)	141 (7%)	102 (5%)

Agrupamiento Simbólico: Algoritmo Cobweb.

El agrupamiento simbólico, como es el caso del algoritmo *Cobweb*, efectúa un análisis cualitativo utilizando aprendizaje incremental, que construye categorías jerárquicas en forma de árbol, donde las hojas representan los segmentos obtenidos. No necesita que se le proporcione el número exacto de clusters deseados, sino que en base a los parámetros *Acuity* y *Cut-off* encuentra el número óptimo [5] [6]. Al modificar el valor del parámetro Cut-off a 0.62 que realiza sobreajuste, con el objetivo de encontrar un número de clusters más manejables; se obtuvieron 6 clusters finales para ambos departamentos, en la Tabla 5 se puede observar de forma resumida los segmentos prioritarios, junto a su cantidad de instancias y porcentajes.

Table 5. Segmentación Algoritmo Cobweb (Clustered Instances)

Dpto. Informática		Dpto. Administración	
2 191 (20%)	3 226 (24%)	2 155 (8%)	3 191 (9%)
5 34 (3%)	6 49 (5%)	4 217 (11%)	6 254 (12%)
7 26 (3%)	4 432 (45%)	7 272 (13%)	8 967 (47%)

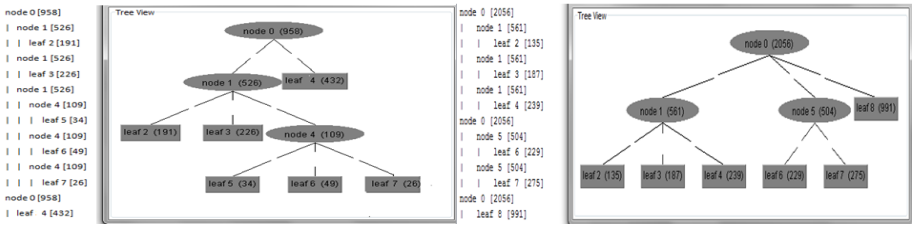

Fig. 3. Segmentación con el Algoritmo Cobweb

En la Figura 3, se expone la estructura de árbol generada por el algoritmo, donde para cada nueva instancia se busca el mejor nodo/segmento (desciende en el árbol). Se coloca la nueva instancia en cada nodo y en uno nuevo, luego mide en cual tiene mayor ganancia de utilidad de categoría (CU); considerando en cada iteración unir los dos mejores nodos evaluados, si esto no resulta beneficioso, considera dividir ese nodo [5] [6]. Se puede observar para el Dpto. de Informática que existen cuatro niveles en general; a partir del nodo inicio se obtienen dos subdivisiones, uno es un seg-

mento/hoja final (leaf 4) el cual abarca cerca al 50% del total de la muestra (432 instancias), y el otro nodo (node 1) vuelve a dividirse en tres segmentos, dos hojas finales (leaf 2 y 3) y un nodo adicional (node 4) que se divide en tres hojas/segmentos finales (leaf 5, 6 y 7). Y para el Dpto. Administración, se puede apreciar la existencia de tres niveles; partiendo del nodo inicial se obtiene tres subdivisiones, uno es un segmento/hoja final (leaf 8) el cual engloba el 47% de la muestra, y otros dos nodos (node 1 y 5) que vuelven a dividirse, el nodo 1 en tres segmentos/ hojas finales (leaf 2, 3 y 4) y el nodo 5 en dos segmentos/hojas finales (leaf 6 y 7).

Para poder ver y analizar en detalle la distribución de instancias, atributos y sus respectivos valores dentro de los clusters generados por el algoritmo *Cobweb* se tuvo que recurrir al algoritmo de clasificación *Naive Bayes*, que obtuvo una precisión del 95,2% para el Dpto. de Informática y 88,3% de instancias clasificadas correctamente.

3.2 Técnicas Predictivas de Minería de Datos.

Selección de Atributos.

El método de evaluación seleccionado es *CfsSubsetEval* y el método de búsqueda es *BestFirst*, con los que se selecciona subconjuntos de atributos de mayor calidad. El método de test para evaluar los atributos es el de *Cross-validation* de 10 folds. Se tienen en cuenta los porcentajes de incidencias de los atributos sobre los valores de los atributos objetos obtenidos con la validación cruzada mayores o iguales al 50% sobre la muestra [7].

En la Tabla 6 se puede observar, para ambos departamentos, que los atributos más relevantes son el promedio general del primer y segundo año en un 100%, con una serie de variantes de influencia de los demás atributos detectados.

Table 6. Promedio General (number of folds (%) attribute)

Dpto. de Informática	Dpto. de Administración
9(90 %) FracC2	10(100 %) AñoAcad
10(100 %) PromGr1	10(100 %) PromGr1
10(100 %) PromGr2	10(100 %) PromGr2
	10(100 %) Loc

Los atributos más significativos son los promedios de materias aprobadas del primer y segundo año, los cuales fueron seleccionados siempre, es decir en cada uno de los 10 folds de validación cruzada (Tabla 7).

Table 7. Promedio de Exámenes Finales Aprobados (number of folds (%) attribute)

Dpto. de Informática	Dpto. de Administración
6(60 %) Apro1	10(100 %) Apro1
10(100 %) PromAp1	10(100 %) PromAp1
10(100 %) PromAp2	10(100 %) PromAp2
7(70 %) EdadIng	50(50 %) EdadIng

En la Tabla 8, se señalan como importantes los atributos “Plan”, “AñoAcad” (año académico), “1°Apr” (materias aprobadas correspondientes al 1° año de la carrera), “Apro2” (materias aprobadas en el 2° año dentro de la universidad) y la “Loc” (localidad) en todos los casos en un 100% (seleccionados en 10 de los 10 folds); pero a diferencia al Dpto. de Informática, no incluye al atributo “1°Reg” (materias regularizadas correspondientes al 1° año de la carrera) como atributo relevante.

Table 8. Condición Final (number of folds (%) attribute)

Dpto. de Informática	Dpto. de Administración
10(100 %) Plan	10(100 %) Plan
10(100%) AñoAcad	10(100 %) AñoAcad
10(100 %) 1°Reg	10(100 %) 1°Apr
10(100 %) 1°Apr	10(100 %) Apro2
10(100 %) Apro2	10(100 %) Loc
10(100 %) Loc	

Clasificación.

Se pretende construir un modelo que permita predecir la categoría de instancias en función de una serie de atributos de entrada, el clasificador aprende un modelo de clasificación de los datos. Uno de los algoritmos de MD más popular y utilizado en multitud de aplicaciones es *C4.5*, que genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de primero en profundidad. La implementación en Weka de este algoritmo se denomina *J48*, el cual presenta los resultados en forma de diagrama y en forma gráfica mediante un árbol. El modo de evaluación seleccionado para los algoritmos de clasificación, es el de validación cruzada, el cual divide varias veces el mismo conjunto de datos en entrenamiento y test (particiones estratificadas); así las particiones de test no solapan. El clasificador final se construye con todos los subconjuntos de datos y la precisión se obtiene del promedio total [7] [8].

Algoritmo Árbol de decisión J48.

Predicción del Promedio General.

Como atributos de entrada, se toman como referencia los atributos seleccionados por la técnica Select attributes como los que más inciden sobre el atributo objeto. Con el algoritmo se obtiene como resultado una serie de condiciones representadas de forma escrita mediante un conjunto de reglas (condiciones del tipo if-else) y gráfica mediante los árboles expuestos en la Figura 4.

Para ambos departamentos, el nodo (condición) inicial representa el promedio general obtenido en primer año (“PromGr1”) y el segundo criterio de clasificación es el promedio general del segundo año (“PromGr2”). La precisión es de un 71,6% de instancias clasificadas correctamente para el Dpto. de Informática y 72,47% para el de Administración.

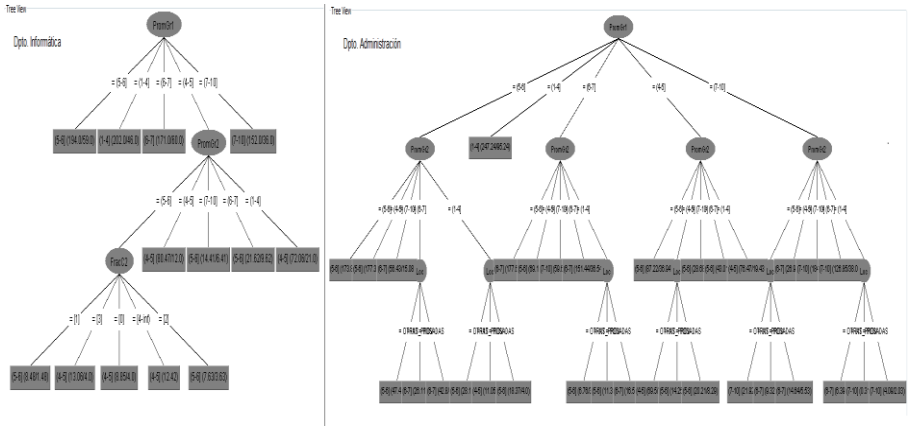

Fig. 4. Árbol de decisión Algoritmo J48 para "PromGralC"

Predicción del Promedio de Exámenes Finales Aprobados.

El algoritmo J48 obtiene como resultados los árboles representados en la Figura 5. La condición inicial para clasificar las instancias en base a los atributos de entrada en relación al promedio de asignaturas aprobadas de la carrera ("PromAproC"), es dicho promedio obtenido en primer año ("PromAp1"), con tres subdivisiones correspondientes a sus valores posibles, el segundo criterio de clasificación para todos los casos, es el promedio de materias aprobadas del segundo año ("PromAp2"). Las precisiones de instancias clasificadas correctamente son del 83% de los casos para el Dpto. de Informática y del 78,94% para el de Administración.

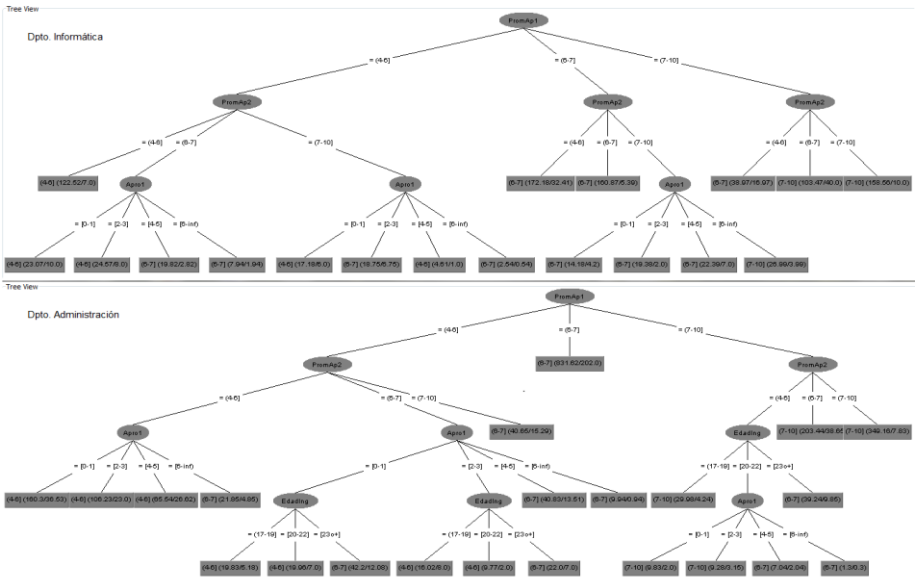

Fig. 5. Árbol de decisión Algoritmo J48 para "PromAproC"

Predicción de la Condición Final.

El ultimo atributo elegido como atributo objeto para determinar su comportamiento, es la condición final, con el que se obtiene una serie de condiciones con una precisión del 61% para el Dpto. de Informática y 65,5% para el de Administración. En la Figura 6 se puede apreciar los árboles de clasificación obtenidos para ambos departamentos; donde se identificó como atributo de mayor incidencia al "AñoAcad" (último año académico en el cual tuvo actividad el alumno), pero el segundo atributo significativo para el Dpto. de Informática es el plan y para el de Administración la localidad ("Loc"), seguido de los demás atributos con variantes según el departamento.

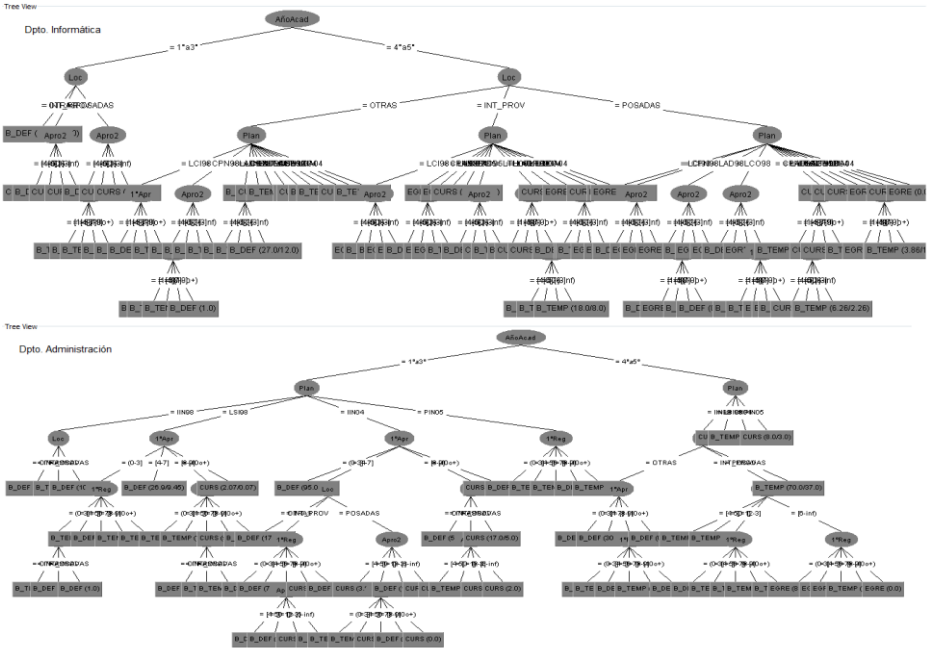

Fig. 6. Árbol de decisión Algoritmo J48 para "Condición"

4 Evaluación e interpretación de resultados.

Como se mencionó, suele surgir la necesidad de volver a comenzar el análisis sobre los datos, variando los mismo, sus distribuciones o incluso con modificaciones sobre objetivos de estudio, dado al carácter dinámico del proceso de KDD. Para esta investigación se ha realizado una gran cantidad de pruebas, consultas a la base de datos, selección de diversos atributos, diferentes discretizaciones de los valores, exhaustivas ejecuciones y evaluaciones de las mismas, entre otras tareas propias del proceso de extracción de conocimiento (KDD).

Al efectuar varias veces el proceso, se han tomado en cuenta ciertos atributos que luego de ser analizados han sido descartados ya sea por su irrelevancia o incompletitud; por ejemplo, asistencias a clases, número de intentos en los exámenes finales correspondientes a todos los años de la carrera, calificaciones de materias del

curso de ingreso, número de intentos en exámenes finales por llamados anuales, títulos obtenidos por alumnos egresados, demora en tiempo para que los mismos egresen, equivalencias, entre otros. Esto lleva a un refinamiento en los resultados, es decir, en el nuevo conocimiento.

Para esta investigación se incluyeron varias carreras, agrupadas en dos departamentos principales, lo cual es un factor diferencial respecto a otros trabajos similares. Esto permitió hacer un análisis comparativo, y a su vez segmentado, considerando las características similares y diferenciales de cada carrera y/o departamento. Este aporte, posibilitaría llevar a cabo acciones diferenciadas entre los departamentos con el fin de realizar ajustes específicos según sus características.

Otro aspecto importante del trabajo es que se consideró el rendimiento académico en dos tramos de la carrera, es decir primero y segundo año como primer tramo y los años superiores como el segundo. Esto permitió reducir niveles de clasificación, obtener resultados agrupados más concretos y ajustar los datos a las estructuras curriculares (en los dos primeros años, las carreras de un mismo departamento, comparten el mayor número de asignaturas comunes).

El análisis de más de un atributo-indicador, representa una contribución adicional, dado que se estudiaron tanto los promedios de asignaturas aprobadas como generales, así como la condición final del alumno que permite apreciar el éxito o fracaso académico desde dos puntos de vistas, las calificaciones obtenidas en exámenes finales, y el grado de éxito de finalización de la carrera (obtención del título), o el fracaso que indica el grado de deserción presente en la universidad. Adicionalmente, se estudiaron varias técnicas de MD, como ser asociación, segmentación, clasificaciones y selección de atributos; siendo lo habitual, centrarse en una o dos técnicas en concreto.

Por último, cabe mencionar que las pruebas son en base a las características reales existentes en la base de datos de la UGD, no estudiado previamente, y mucho menos aun preparado para este tipo de experimentos/investigaciones.

Respecto a los resultados por departamento.

Un análisis general de los datos, permitió detectar ciertas diferencias entre departamentos como ser: el año académico predominante en la muestra correspondiente al Dpto. de Informática es entre primer y tercer año, mientras que para el de Administración entre cuarto y quinto año. También, las edades al ingreso se diferencian, ya que en Informática ingresan en su mayoría jóvenes recién egresados del secundario con edades entre 17 y 19 años y en cambio, en Administración ingresan con edades que promedian los 23 años, esto puede estar relacionado con carreras previas o equivalencias de otras universidades, o simplemente inicio tardío a estudios universitarios. Otro aspecto detectado, es que los alumnos de Informática cursan más asignaturas que los de Administración en los dos primeros años de sus carreras (por las características de los planes de estudio), pero aprueban prácticamente la misma cantidad de materias y los fracasos de cursado (condición libre o re-cursante) en primer y segundo año, son mayores que los obtenidos por los de Administración, en este último la cantidad es muy baja. Adicionalmente, se detectó un comportamiento sistemático en el rendimiento por parte de los alumnos de Informática, dado que al pasar de primero a segundo año y a lo largo de sus carreras, los promedios fueron relativamente constantes,

en cambio para los de Administración, estos disminuyeron en el segundo año de sus carreras.

En el Dpto. de Informática, la cantidad de asignaturas aprobadas del primer año marca tendencia sobre el resto de la carrera, es decir si llegan o no al último año de la carrera, con un valor de confianza aproximado del 80%. Otra cuestión visible aquí se relaciona con la condición final del alumno, donde, del total de alumnos en condición de baja temporal, el 74% de ellos están en cuarto o quinto año de la carrera, situación que no se destacó para el otro departamento, lo cual puede estar relacionado con la tesis de finalización de carrera, que por lo general demoran más que las de administración.

Respecto a los patrones y las relaciones entre los datos seleccionados para la investigación.

Los promedios del primer y segundo año de la carrera marcan una tendencia e influyen de forma directamente proporcional (en la mayoría de los casos) sobre el promedio de asignaturas aprobadas y generales de la carrera. Dicha tendencia se vio claramente reflejada mediante la técnica de asociación, de selección de atributos y la de clasificación, con porcentajes de precisión superiores al 70%, dependiendo de la combinación con otros atributos como localidad, cantidad de fracasos en materias cursadas, número de materias aprobadas, entre otros.

Por medio del algoritmo de asociación *Apriori*, se demostró y destacó que la mayoría de las reglas obtenidas en las pruebas, hacen referencia a que la condición mayoritaria de bajas definitivas ("B_DEF") ocurren entre el primer y tercer año de la carrera; donde los porcentajes de confianza varían entre un 70% y 90% dependiendo de la combinación con otros atributos, como ser la localidad de origen "Otras", entre 1 y 3 asignaturas aprobadas del primer año de la carrera, y una o ninguna materia aprobada en el segundo año calendario dentro de la institución. Las reglas obtenidas en relación al año académico, demuestran una estrecha relación entre los promedios de aprobadas y generales, y la cantidad de alumnos que llegan al cuarto o quinto año de la carrera, donde los coeficientes de confianza obtenidos varían alrededor del 70% y 80% de los casos para ambos departamentos.

Conclusiones.

Las herramientas de MD brindan resultados que deben ser interpretados y traducidos a diagnósticos y consecuencias del ámbito real (en este caso la institución educativa). Esto implica que los resultados de la aplicación de las técnicas se han utilizado para explicar parte del comportamiento de la situación en cuanto a desempeño académico de los alumnos, pero no se agota en el análisis realizado en este informe. Como así también, las posibles consecuencias y acciones tendientes a la toma de decisiones específicas, está sujeta a consideraciones de otros integrantes del cuerpo académico de una institución educativa.

Corresponde observar que la información académica utilizada para las pruebas realizadas proviene de una institución educativa en crecimiento y adaptación, es decir que tanto las estructuras organizativas como los planes de estudio y las metodologías utilizadas, inclusive el cuerpo docente, tuvieron cambios significativos durante el período del cual se tomaron los datos. En particular, los cambios en los planes de estudios afectaron tanto en el desarrollo de las carreras como al proceso de análisis aplicado en el trabajo y por tanto tuvo efectos en los niveles de precisión de los resultados obtenidos.

Tal como se aprecia en otros trabajos similares, en el proceso KDD la preparación y acondicionamiento de los datos es la etapa más extensa y a la vez fundamental porque en gran medida los resultados posteriores dependen de ésta.

En las etapas intermedias, es crítico llevar a cabo análisis e interpretaciones de resultados parciales, ya que a partir de éstos se retoma el proceso y continúa la depuración y refinamiento del conocimiento extraído. Se recomienda utilizar distintas variantes, todas las que sean posibles.

En el presente trabajo se llevó a cabo un análisis comparativo y segmentado entre departamentos académicos, lo cual permitió identificar:

- Características similares y diferenciales, que requieren acciones diferenciadas entre distintos departamentos, con el fin de realizar ajustes específicos según sus características.
- Necesidades diferentes en cuanto a estrategias de apoyo por edades, análisis de contenidos, prácticas y metodologías de los espacios curriculares, acciones de retención y apoyo a alumnos de los primeros años y distintas estrategias de motivación.

Referencias Bibliográficas.

1. C. Pérez López: “Minería de Datos – Técnicas y Herramientas”, 2007, Editorial Thomson, 1a Edición, Madrid España.
2. P. Britos, A. Hossian, R. García Martínez, E. Sierra: “Minería de Datos basada en Sistemas Inteligentes”, 2005, Editorial Nueva Librería, 1a Edición. Buenos Aires Argentina.
3. J. Hernández Orallo, M.J. Ramírez Quintana y C. Ferri Ramírez: “Introducción a la Minería de Datos”; 2004, Editorial Pearson, 1a Edición, Madrid España.
4. U.M. Fayyad, G. Piatetsky Shapiro, O. Smyth, R. Uhturudsamy: “Advances in Knowledge Discovery & Data Mining”; 1996, Editorial MIT Press, Cumberland.
5. M. Garre, J. Cuadrado, M. Sicilia, D Rodríguez, R. Rejas: “Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software”, Revista Española de Innovación - Calidad e Ingeniería del Software, 2007, Vol.3 No. 1.
6. A. Moreno, E. Armengol, E, J. Béjar, L. Belanche, U. Cortés, R. Gavaldá, J.M. Gimeno B. Lopez, M. Martín, M. Sánchez: “Aprendizaje automático”, 1994, Editorial UPC, 1a Edition, Barcelona España.
7. I. Witten, E. Frank: “Data Mining: Practical Machine Learning Tools and Techniques”, 2005, Editorial Morgan Kaufmann, 2d Edition. San Francisco EEUU.
8. J.M. Molinas Lopez, J. García Herrero: “Técnicas de Análisis de Datos – Aplicaciones prácticas utilizando Microsoft Excel y Weka”, Universidad Carlos III de Madrid España. [En línea] http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf.