Trabajo de tesis para optar el título de:
MAGISTER EN INGENIERÍA DE SISTEMAS Y COMPUTACION

Inteligencia de negocios aplicada al problema de adquisición y retención de estudiantes en una universidad privada

RAFAEL CASTILLO SANTOS
Facultad de ingeniería de Sistemas y Computación

2006

TABLA DE CONTENIDO

TABLA DE FIGURAS

LISTA DE TABLAS

Agradecimientos.

Quiero dar mis sinceros agradecimientos a todas aquellas personas que hicieron posible la culminación de este trabajo, especialmente a la Universidad objeto de estudio por la ayuda económica brindada y que por medio de Alexa Corena me brindo los datos para la realización y aplicación de la herramienta; a mi asesor Doctor José Abásolo Prieto quien con sus conocimientos me dio las guías necesarias para el adecuado planteamiento del trabajo; al profesor Germán Bravo por sus comentarios y mejoras dadas como jurado de tesis; Al Doctor José María Álvarez Manrique también por sus guías y correcciones dadas como jurado y como persona conocedora en profundidad del tema de deserción universitaria; a mi familia por soportar todo el tiempo que no les pude dedicar y a todos los compañeros de estudio con quienes compartí y a quienes me dieron el apoyo para culminar mis estudios.

1 INTRODUCCIÓN

La educación como factor dinámico del desarrollo de una sociedad esta siendo cuestionada hoy debido a los resultados logrados en cada uno de los estamentos que lo constituyen: educación preescolar, primaria, secundaria, universitaria y posgradual. Los primeros tres tienen una amplia participación del estado y de alguna manera se considera obligatoria, pero la educación superior (universitaria y posgradual) no tiene esa connotación razón por la cual es de libre elección por quienes quieran estar o no en ella. Sin embargo justamente es la educación superior la que proporciona el motor de desarrollo de un país en sus áreas científica y tecnológica. Hace unos pocos años, el que un estudiante de educación superior continuara o no sus estudios no se consideraba un problema, pero hoy en día se considera como un índice de pobres resultados o ineficiencia de las instituciones universitarias, sean estas de carácter público o privado. La situación tiene ahora la atención del estado debido a que se considera un problema que va en aumento y al cual hay que darle el tratamiento y soluciones necesarios, tal como lo expresa Javier Botero, viceministro de Educación "Si de cada casi dos estudiantes uno no termina con éxito, estamos siendo ineficientes en todo sentido en el uso de los recursos y de la infraestructura, y en los esfuerzos. No basta con haber incrementando el número de estudiantes que presentan el Icfes y que ingresan a la educación superior, que pasó del 38% en 2002 al 66% en 2006, si dejamos salir a la mitad".

La deserción estudiantil universitaria no es un problema nuevo, y ha tenido diferentes formas de tratarlo dependiendo de cada institución, pero hasta ahora no hay una aplicación informática que haga uso de las herramientas de inteligencia de negocios, concretamente de minería de datos. Con base en lo anterior se presenta este trabajo para mostrar las bondades de la aplicación de la inteligencia de negocios en la solución al problema de adquisición y retención de estudiantes en una universidad privada. Se hace énfasis en la deserción universitaria debido a que es un problema crítico que se presenta actualmente en la mayoría de las universidades no solamente de Colombia, sino de la mayoría de países latinoamericanos y adicionalmente por la necesidad que tienen las universidades de carácter privado de competir por la vinculación de nuevos estudiantes, retener a los que ya están y evitar las deserciones, situación que las condiciona cada vez más a competir y por lo tanto permanecer o no en el mercado.

El proceso realizado se llevó a cabo como proyecto de minería de datos orientado a establecer cómo las estrategias orientadas a fortalecer las condiciones de retención y solucionar las relacionadas con la deserción, mejoran las condiciones de la organización y su situación económica, aumentando los índices de participación en el mercado.

Aplicaciones informáticas sobre el tema ya han sido realizadas en diferentes instituciones, pero la mayoría son de tipo estadístico en cuanto a saber por ejemplo, cuántos estudiantes ingresaron, de ellos cuántos son hombres y cuántos

mujeres, que edad es la que predomina al ingreso o cuál al terminar sus estudios, etc. Entre estos proyectos hay uno que se destaca y esta a disposición de la comunidad universitaria, es el desarrollado por el Ministerio de Educación Nacional en convenio con el Centro de Estudios Económicos de la Universidad de los Andes (CEDE), llamado SPADIES (Sistema de Prevención y Atención de la Deserción en las Instituciones de Educación Superior). Los datos fueron proporcionados por cada una de las universidades participantes en el proyecto y con base en ella les permite obtener diferentes consultas de sus propios datos y comparase en diferentes aspectos con todas las demás tanto a nivel local como a nivel nacional. Determina cuál es el riesgo que el estudiante tiene de no terminar sus estudios pero no el perfil del estudiante antes de realizar su ingreso a la institución, lo cual sí permitirá hacer la herramienta planteada como prototipo de aplicación del presente trabajo.

Los beneficios del proyecto se dan en la medida que se dé solución a las dificultades antes mencionadas, utilizando la herramienta prototipo, al predecir cuántos de los estudiantes que ingresan en un período lectivo desertarán o cuántos continuarán sus estudios y de qué manera irán desertando hasta la finalización de la carrera elegida. Igualmente a medida que se apliquen las estrategias orientadas a enfrentar el problema, poder ver sus efectos. Para los estudiantes el proyecto dará las facilidades necesarias de determinar con suficiente anticipación, cuáles serán las asignaturas que se ofrecerán para un período académico, permitiéndoles orientar mejor su plan de estudios y perfil profesional.

Con base en el estudio y conocimiento de la organización en las situaciones descritas, se desarrolló un prototipo con modelos de minería de datos tanto para entender el problema y plantear soluciones a la retención, como para el retiro (o deserción) voluntaria e involuntaria de estudiantes de la universidad. Como fuentes de información se tomaron los datos históricos y se identificaron las variables. Del análisis de cada una de estas variables se determinaron las áreas dónde se es fuerte y aquellas dónde se tienen debilidades.

El aporte que el presente trabajo hará a la comunidad es de tipo práctico/técnico, pues resulta novedoso el hecho de aplicar minería de datos a problemas que tienen las organizaciones de carácter educativo (tanto universidades, colegios, etc) relacionados con su principal fuente de ingresos: los estudiantes.

El proyecto por lo tanto constituye una ventaja competitiva que la universidad objeto de estudio, tendrá para mejorar la atención de los estudiantes, ofrecer oportunamente programas académicos más adecuados y por tanto obtener mejores ingresos en sus resultados. Desde el punto de vista de tema de trabajo para la maestría permite confrontar la teoría que existe sobre minería de datos, con el ambiente de aplicación en una institución universitaria y obtener los resultados, que posiblemente puedan ser aplicados en otras instituciones de similares características.

Los siguientes son aspectos que constituyen fundamento suficiente para sostener la propuesta de implementar procesos de minería de datos en la Universidad:

El crecimiento de los datos:

- El número de registros u objetos.
- El número de campos o atributos de un objeto.

Herramientas para la toma de decisiones en la obtención de nuevos conocimientos: Hoy día, es natural utilizar técnicas computacionales que ayuden a descubrir patrones y estructuras en grandes almacenamiento de datos. El descubrimiento del conocimiento (*Knowlegde Discovery in Databases, KDD*, en inglés) es una herramienta para manejar el problema de sobrecarga de datos con la exigencia de responder retrospectivamente en el tiempo, brindando datos en múltiples niveles de análisis. En el presente trabajo, la tecnología de minería de datos complementa la infraestructura actual aportando respuestas a muchas de las preguntas y situaciones concretas de una institución universitaria.

Soporte computacional que favorece la minería de datos: La combinación bodegas de datos – minería de datos, aunque no necesaria, es la fórmula más utilizada para la implementación de procesos de descubrimiento del conocimiento.

2 OBJETIVO GENERAL

Buscar soluciones a los problemas de retención y deserción universitaria, utilizando técnicas de Minería de datos.

3 OBJETIVOS ESPECIFICOS

1. Estudiar el problema de retención y deserción de estudiantes en un contexto universitario.

2. Plantear soluciones basadas en técnicas de minería de datos, para resolver esos problemas.

3. Hacer una prueba piloto en un caso real de una Universidad Colombiana.

4 JUSTIFICACIÓN

Durante el tiempo transcurrido desde la fundación de las universidades, la toma de decisiones la han realizado los directivos fundamentalmente con base en la experiencia, después de realizar largos y tediosos análisis del comportamiento de la información, obtenida ésta, con herramientas que aunque facilitan el proceso, lo hacen de manera muy lenta y luego de interminables cruces.

Actualmente el proceso de acreditación para la calidad de la educación superior, condiciona a las universidades y establece algunos lineamientos orientados a mejorar la calidad. Algunos de ellos están relacionados con calidad académica, bienestar universitario, contratación de docentes y controles para la deserción universitaria. Para el manejo de estas variables, es importante contar con un sistema de información que permita determinar comportamientos, hallar proyecciones y relacionar diferentes variables. Por lo tanto la minería de datos sobre información histórica se constituye en una herramienta útil al permitir facilidades de generación de reportes y descubrir y entender patrones ocultos en bodegas de datos. Estos reportes y patrones permiten entender comportamientos y tendencias con muy alta precisión por medio del uso de algoritmos supervisados o no supervisados. ¿El Resultado? Las instituciones universitarias podrán de una manera más adecuada realizar todos sus procesos de proyección, planeación y gestión de los recursos.

Por otra parte, el problema de deserción y retención de estudiantes se ve hoy como un signo de ineficiencia y como un gran costo para el país, para los estudiantes y para cada una de las instituciones de educación superior. La deserción universitaria pasó de ser una cifra programada que estaba en los presupuestos de las universidades, a convertirse en un problema que hay que entender para poder combatirlo.

Aunque la deserción es un fenómeno antiguo, la Universidad en estudio, maneja los datos con metodologías propias y de forma global, pues no hay un instrumento que permita hacerle un seguimiento a cada estudiante matriculado y determinar los factores que ponen en riesgo su continuidad, dé las señales de alerta a tiempo para asegurar su permanencia en la institución.

5 EL PROBLEMA DE RETENCIÓN Y DESERCIÓN DE ESTUDIANTES EN UNA UNIVERSIDAD PRIVADA.

Las condiciones por la cuáles una institución universitaria o de cualquier otro tipo no alcance sus metas de rentabilidad, de ocupación de sus servicios, etc, es debido a que sus clientes, en este caso estudiantes, no lleguen en la cantidad esperada, que se retiren antes de culminar la totalidad del programa o que no se sientan satisfechos con la calidad de los servicios obtenidos.

Tampoco hay suficiente conocimiento por parte de los organismos académicos - administrativos universitarios del comportamiento de los estudiantes con las razones por la cuáles continúan su programa académico o se retiran del mismo. Ni hay estudios institucionales completos de mercadeo para ofrecer planes y programas que satisfagan las necesidades y situaciones de dificultad económica de los estudiantes.

Desde el punto de vista competitivo la oferta de educación superior es bastante alta comparada con la demanda que de la misma están haciendo los estudiantes de bachillerato[1], debido por un lado a dificultades de tipo académico y por otro a la promoción de educación no formal o educación técnica a precios mucho mas bajos y duración de estudios de menor duración. Lo anterior unido a la mayor demanda laboral de personas con capacidades técnicas y los bajos salarios con los cuales se remunera.

Así, para el caso de los estratos 1, 2 y 3 se ve una tendencia significativa a la deserción debido fundamentalmente a que la edad de estos estudiantes oscila entre 17 y 20 años, en la cual ya están aptos para el ingreso a la población económicamente productiva; a la disminución de los ingresos familiares y por otra parte a la mayor valoración del trabajo material antes que el intelectual.

Quizá el mayor impacto de estas estadísticas, tal como lo muestra la figura 1, está en el ingreso familiar, el cuál para ingresos de estratos 5, cerca del 99% de los estudiantes ingresan y permanecen en la educación superior, mientras que para estrato 4 la cifra es de apenas el 37% y 18% para estrato 3.

[1] Entre 1993 y 2000 la matrícula de secundaria aumentó cinco veces más que la población en edad de cursar ese nivel y la de preescolar incrementó en 11%. Como un todo, la oferta educativa creció en 2'390.000 cupos, mientras que la población lo hizo en un 1'500.000 personas [10]

Quintil de ingreso per cápita	Cobertura de educación Superior		Diferencia TB-TN	Extraedad	
	Tasa bruta de cobertura	Tasa neta De cobertura		<18 años	>24 años
1	3.6	2.3	1.4	0.3	1.1
2	9.7	6.0	3.6	0.7	3.0
3	18.0	10.1	7.9	1.2	6.7
4	37.8	18.6	19.2	1.9	17.3
5	99.0	53.1	45.9	2.9	43.0
Total	**30.0**	**16.1**	**13.9**	**1.3**	**12.6**

Fuente: Cálculos de la Misión Social del DNP con base en Encuesta de Calidad de Vida, 1997.

Figura 1. Cuadro comparativo de ingreso a la educación superior con relación a los ingresos familiares [6].

Sin embargo un factor positivo en esta dinámica lo muestran las estadísticas de empleo. Tal como lo muestra el cuadro de la figura 2, Los salarios de los ocupados con educación superior crecieron entre 1991 y 1998 un 26% en las zonas urbanas, mientras que para las personas con sólo estudios básicos de primaria decrecieron en un 18%. El aumento de salarios ha favorecido a los que tienen educación superior y la ocupación para personas con secundaria y universitaria aumentó en un 18% y 41% respectivamente, mientras que la tasa de empleo de las personas sin educación o con educación primaria bajó. Una condición que afecta el ingreso a los niveles salariales altos o al ingreso a determinados cargos es la exigencia de un mayor nivel educativo (figura 3.)

Zona	1991		1994		1998	
	Ocupados	Salario medio	Ocupados	Salario medio	Ocupados	Salario medio
Urbano						
Ninguna		0.84	195,468	1.13	267.692	0.90
Primaria	2,539,926	1.53	2,568,186	1.59	2,436,034	1.28
Secundaria	3,743,908	1.76	4,323,085	2.29	4,497,936	1.78
Superior	1,295,356	3.83	1,614,467	5.77	1,961,838	4.97
Subtotal*	7,869,025	2.00	8,723,186	2.70	9,201,094	2.30
Rural						
Ninguna		0.89	816,626	0.88	1,107,505	0.92
Primaria	3,688,823	1.21	3,554,629	1.03	3,324,455	0.93
Secundaria	1,183,237	2.08	1,282,235	1.65	1,223,566	1.60
Superior		4.26	149,902	3.64	213.110	3.84
Subtotal*	5,887,266	1.39	5,803,392	1.21	5,878,769	1.18
Total						
Ninguna	1,160,038	0.88	1,012,094	0.93	1,375,197	0.91
Primaria	6,228,749	1.34	6,122,814	1.27	5,760,489	1.08
Secundaria	4,927,145	1.84	5,605,320	2.14	5,721,502	1.74
Superior	1,402,644	3.87	1,764,369	5.59	2,174,948	4.86
Subtotal*	13,756,291	1.74	14,526,578	2.11	15,079,863	1.86

* El número restante para alcanzar el total corresponde a las personas que no informan.

Fuente: DANE-Encuesta de Hogares para septiembre de cada año.

Figura 2. Empleo y salarios según nivel educativo. [6]

Nivel educativo/ ocupación	1978 %	1991 %	1993 %	1995 %	1997 %	1999 %
Alguna universitaria y más						
Profesionales, técnicos y directivos	27.9	70.9	72.1	75.5	77.6	79.4
Personal administrativo	2.3	26.1	26.2	28.7	30.4	30.2
Comerciantes, vendedores	0.6	8.0	9.7	9.9	10.9	10.1
Trabajadores de los servicios	0.3	2.1	2.4	2.8	4.0	3.9
Trabajadores agrícolas	0.8	0.6	0.5	0.7	0.9	0.9
Trabajadores operarios no agrícolas	0.4	3.1	3.6	3.7	3.8	3.9

Fuente: Cálculos Departamento Nacional de Planeación-UDS-DIOGS con base en Dane, Encuesta Nacional de Hogares, Septiembre.

Figura 3. Población ocupada, para educación universitaria o superior. [6]

Desde otro contexto el problema puede ser visto teniendo en cuenta las consideraciones presentadas por el diario la Prensa en su edición de Junio 20 de 2005 [7]. En este artículo se comenta que el porcentaje de estudiantes que no terminan su carrera universitaria esta alrededor del 52%., es decir de cada dos estudiantes que ingresan a la educación superior solamente uno termina. Entre las razones que se manifiestan para la deserción están:

- Razones económicas.
- Mala calidad de los programas. “51 programas están a punto de ser cerrados por el incumplimiento de requisitos de calidad” [7].
- Dificultades en el aprendizaje y en la lectura.
- Falta de información sobre la vida universitaria.
- Ausencia de una orientación vocacional.

La mayoría de las universidades apenas están enfrentando el problema, razón por la cual no hay aún planes concretos que faciliten la permanencia de los estudiantes en el claustro universitario, que hagan promoción de los egresados a su actividad laboral y establezcan convenios con las organizaciones para permitir que los estudiantes hagan pasantias profesionales.

5.1 IDENTIFICACIÓN DEL PROBLEMA.

La deserción académica es un fenómeno que afecta de manera general a las instituciones educativas del mundo, pero las causas que la generan presentan particularidades de un país a otro y, por tanto, de una institución a otra, aunque en últimas responde a toda una problemática de crisis de las instituciones de educación superior, de la que no podía estar excluida la Universidad en estudio.
El problema se manifiesta por la pérdida de estudiantes que cursan regularmente sus estudios, tal como se muestra en la figura 4 para la cohorte del 2001-1 al 2006-1, y en el ingreso insuficiente o no ingreso de nuevos estudiantes, reflejado en el figura 5 para el período 2001-1 al 2006-1.

En cuanto al problema de deserción, este se constituye como tal, pues como se observa en la figura 4, la deserción para la Universidad en estudio, está muy próxima al 70% (de cada 10 estudiantes que ingresan 3 terminan) la cual esta lejos de la media normal de otras universidades, donde el índice esta alrededor del 50 %[2] (de cada 10 estudiantes 5 terminan). El problema se hace más evidente cuando como se muestra en la figura 5, la cantidad de estudiantes que ingresan (primíparos) se mantiene estable o con tendencia a la baja. Esto origina dificultades de elaborar el presupuesto de los recursos tanto locativos cómo de planta de personal docente y administrativo, originando mal uso de salones, asignación de carga académica, planeación de cursos de asignaturas obligatorias y electivas, cursos de posgrado, etc que conllevan a la disminución real de la rentabilidad.

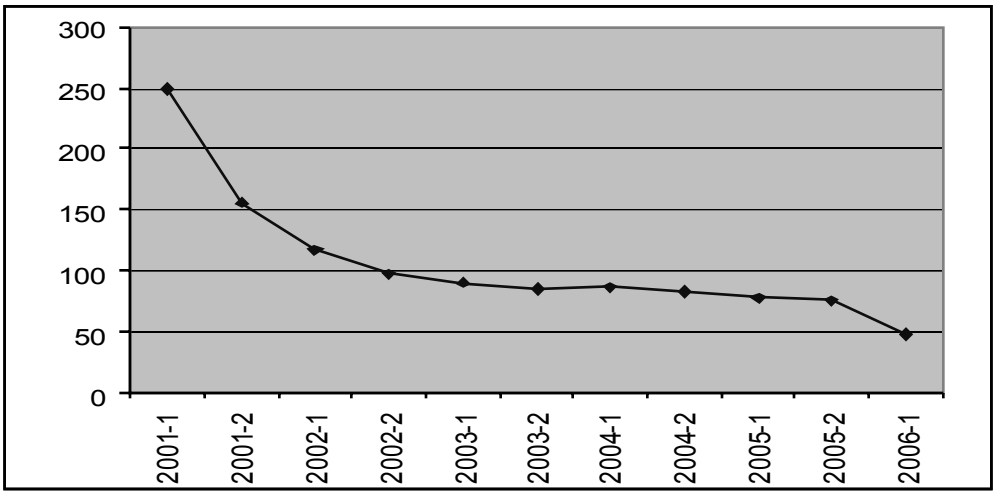

Figura 4. Deserción de estudiantes para la cohorte del 2001-1 al 2006-1.

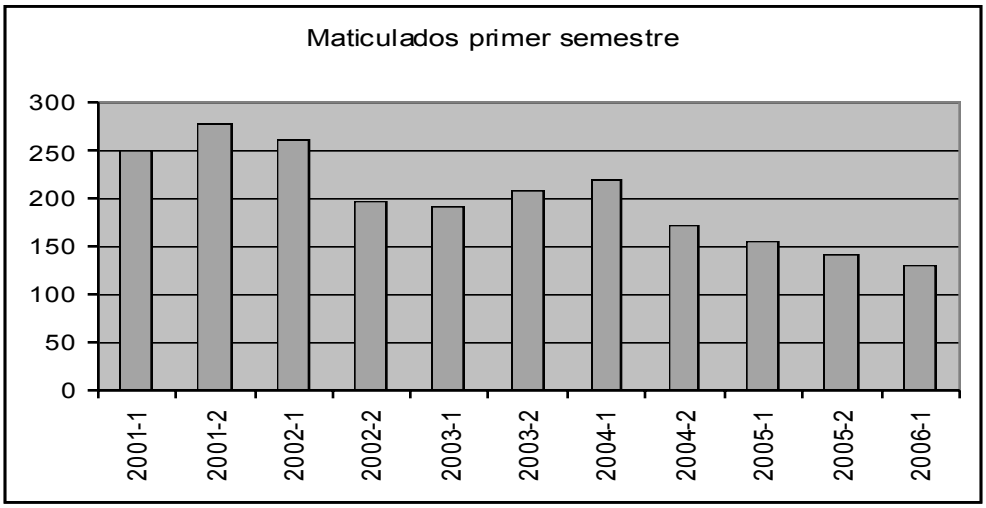

Figura 5. Ingreso de nuevos estudiantes para el periodo del 2001-1 al 2006-1.

[2] Fuente: ministerio de educación nacional SNIES.[14]

5.2 DEFINICION DEL PROBLEMA

La retención y la deserción universitaria pueden entenderse como las medidas porcentuales relacionadas con la cantidad de estudiantes que permanecen en la universidad y terminan su carrera y la cantidad de estudiantes que por alguna razón se retiran no culminando la totalidad de sus semestres y asignaturas respectivamente.[3]

5.3 CLASIFICACIÓN DE LA DESERCIÓN UNIVERSITARIA

No voluntaria. Se presenta cuando es la universidad la que toma la decisión de no permitir que el estudiante continúe sus estudios en la institución. Las razones pueden ser entre otras: Cancelación de la matricula académica por fraude, bajo rendimiento académico (el promedio por debajo del mínimo permitido) en períodos consecutivos y en general el incumplimiento de alguna de las normas de comportamiento establecidas en el reglamento estudiantil. Esta es quizá, la de más fácil detección, pues se tienen los datos precisos de cuándo se lleva a cabo el evento y se puede determinar las razones por las que ocurrió. La universidad del llano [4] plantea que la deserción no voluntaria determinada por el bajo rendimiento académico debe ser considerada como "Mortalidad académica" y no como deserción propiamente dicha.

Voluntaria. Se presenta cuando el estudiante es quien toma la decisión de no continuar con sus estudios en la institución. Las razones para que se dé este evento pueden ser económicas, personales, sociales, ambiente académico o físico, mala preparación en sus estudios secundarios, entre las que mas influyen. Es difícil su detección oportuna, porque en la mayoría de la veces no se tiene el momento cuando ocurrió, (tan solo se detecta al final de un periodo lectivo, o al comenzar el siguiente) pero no el día en que ocurrió, a no ser porque el estudiante haga el comunicado expreso de su deseo de no continuar. La entrevista a posteriori que el estudiante dé de lo ocurrido es de muy difícil consecución, además que tampoco es fácil la ubicación del estudiante una vez se retira y que éste esté dispuesto a suministrarla. Sin embargo con base en los datos históricos del estudiante se puede obtener un perfil del mismo y por lo tanto a través de algoritmos de minería de datos hacer la proyección, clasificación e incluso predicción de los estudiantes con perfiles similares para tomar las medidas oportunas conducentes a plantear estrategias de acuerdo al por qué probablemente un estudiante desertará.

---

[3] Estadísticas e indicadores de la universidad Nacional de Colombia [6]

[4] Malagón Escobar, Luz Miriam, Calderón Cañón, Cesar Augusto y Soto Hernández, Edwin Leonardo. Estudio de la deserción estudiantil de los programas de pregrado de la universidad de los llanos (1998-2004). Villavicencio, Meta Enero de 2006. [13] [32]

Sobre las clases de deserción en educación, se mencionan las siguientes, no excluyentes entre si[5]:

- Deserción total: abandono definitivo de la formación académica individual.
- Deserción discriminada por causas: según la causa de la decisión.
- Deserción por Facultad (Escuela o Departamento): cambio facultad - facultad.
- Deserción por programa: cambio de programa en una misma facultad.
- Deserción a primer semestre de carrera: por inadecuada adaptación a la vida universitaria.
- Deserción acumulada: sumatoria de deserciones en una institución.

Adicionalmente, se involucran en el fenómeno de la deserción, como actores relevantes, no sólo a los desertores, sino también a padres de familia de desertores, ex-compañeros de estudio, profesores, directivas y administradores académicos.

En la universidad del llano, según un estudio[6] entregado en Enero de 2006 que comprende los años 1998 a 2004, plantea la deserción universitaria en cuanto al tiempo y al espacio. En el tiempo la clasifica en:

- Deserción precoz, cuando el estudiante admitido no se matricula.
- Deserción temprana, cuando el estudiante se retira en los primeros cuatro semestres.
- Deserción tardía cuando el estudiante abandona después del quinto semestre.

Con relación al espacio, el mismo estudio determina que es Interna o del programa (cambio de carrera) y del sistema educativo.

En cuánto a indicadores y metodología para calcular la deserción, autores como Osorio, Jaramillo y Jaramillo[7] han tenido en cuenta algunos como[8]:

- Índices de deserción semestral: relación entre el número total de alumnos desertores del programa i en el período t y el número total de estudiantes matriculados en dicho programa para el mismo período.

- Índices de deserción por cohorte: diferencia entre el número de estudiantes que ingresan a la cohorte c en el período t y la cantidad de ellos que se matriculan en el período t +1.

- Índices de deserción promedia por nivel: promedio simple de los índices de deserción por semestre calendario, calculados como el número total de

[5] Universidad Nacional de Colombia. Convenio 107/2002 UN-ICFES. Documento Sobre estado del Arte [15]
[6] Ibid, Malagon Escobar, Luz Miriam.
[7] Osorio, Ana; Jaramillo, Catalina; Jaramillo, Alberto**.** Deserción estudiantil en los programas de pregrado 1995-1998. Oficina de Planeación Integral. Universidad EAFIT, Medellín, 1999. www.eafit.edu.co/planeacion/final.html
[8] Ibid. UN-ICFES. Convenio 107/2002

desertores de cada nivel sobre el total de matriculados en dicho nivel del programa i.

- Tasa ponderada de deserción por nivel: muestra la expectativa de deserción para el programa i. Se calcula ponderando la tasa de deserción con el promedio de la distribución de la población matriculada, en los once semestres de duración de la carrera.

Para el estudio de la transferencia interna (traslado de carrera) como parte de la deserción no académica, en la misma investigación se consideraron dos índices:

- Índice de Recepción: relación entre alumnos recibidos y alumnos cedidos por determinado programa académico. Con este índice se espera hacer una clasificación de las carreras como "dadoras" y "receptoras" de alumnos para identificar hacia qué programas académicos se inclinan los estudiantes desertores de determinado programa y las causas de esa elección.

- Índice de Participación: este índice muestra la participación porcentual de una carrera determinada en el movimiento total de cambios de un grupo de carreras.

Adicionalmente se consideraron las siguientes variables:

- Deserción: abandono que el alumno hace del programa antes de su culminación, conforme al reglamento académico, bien sea por razones disciplinarias -denominada deserción académica (DA)- o por motivos personales -deserción no académica (DNA) o retiro voluntario. Se mide por la diferencia entre la matrícula inicial y la final en un mismo período considerado.

- Deserción Académica: abandono del aula por razones estrictamente académicas

- Deserción No Académica: abandono voluntario que el alumno hace de las actividades académicas a lo largo del programa y cuyas causas pueden ser de tipo exógeno o endógeno a la Institución.

- Episodio de Deserción: cancelación de la matrícula de un estudiante, bien sea por decisión de la Institución (DA) o del alumno mismo (DNA). Así, un estudiante con dos cancelaciones de matrícula en su historia académica genera dos episodios de deserción.

- Preincidente: estudiante que registró en su historia académica más de un episodio de deserción durante el período estudiado.

- Nivel de Deserción: semestre académico en el cual el estudiante abandona sus estudios, bien sea voluntaria o forzosamente. Para aquellos estudiantes que en

el momento de retiro estaban cursando materias de varios semestres, el nivel fue estimado por el número de créditos aprobados.

- Semestre de Retiro: todos los semestres calendario dentro del período 1995-1 y 1998-2 de los cuales desertaron los estudiantes por razones académicas y no académicas.

- Reintegro - Aspirante en Reingreso - Aspirante a Transferencia Externa - Aspirante a Reingreso con Grado Previo - Aspirante a dos o más Programas Académicos en la misma Universidad.

- Estado: Activo: Estudiante que tiene matrícula vigente en cualquier programa académico de pregrado de la Institución. Inactivo: Estudiante no vinculado actualmente a la Institución.

6 METODOLOGÍA.

Las técnicas de minería de datos, permiten la clasificación de datos según un contexto o comportamiento, por ejemplo de ingresos o retiros de estudiantes. Adicionalmente permite obtener comportamientos por medio del análisis secuencial; permite la totalización de datos y la visualización de los mismos por medio de diferentes representaciones (árboles de decisión, funciones lineales o no lineales, modelos de probabilidad, por mencionar solamente algunos).

Los resultados a tener de la aplicación de un proceso a través de minería de datos, son uno o varios modelos de decisión que deben haber sido probados suficientemente, para que puedan ser aplicados por uno o varios departamentos o secciones de la universidad. Estos modelos se comportan como agentes inteligentes, dónde la inteligencia está dada por la aplicación de las políticas y reglas de la universidad. Dos modelos usados frecuentemente en minería de datos son el descriptivo y el predictivo.

Modelamiento descriptivo. Este modelamiento presenta las características fundamentales de los datos a ser estudiados, sin que para ello se creen trabas a la cantidad de los datos. Como resultado se obtiene un entendimiento más concreto de las relaciones de los datos y sus correspondientes estructuras. El modelo se considera que es Generativo, debido a que los datos resultantes tendrán las mismas características de los datos reales de los cuales se generan. [12]

Modelamiento predictivo. El objetivo principal al aplicar un modelo predictivo, es el de encontrar el valor no conocido de una variable, dados los valores o comportamientos de otras variables. Por ejemplo en el caso planteado como titulo de este trabajo, el de encontrar cuántos estudiantes, nuevos en la universidad desertarán de su carrera y de la universidad. Este tipo de modelamiento se puede presentar como la relación un grupo de valores en un vector y sus resultados esperados, obtenidos por medio de una función $y = f(x;\theta)$ donde x son los valores medidos y θ los parámetros del modelo.

La propuesta metodológica a llevarse a cabo tiene como soporte la metodología propuesta por CRISP-DM (Cross-Industry Standard Process for Data Mining)[9] y la planteada por Berry- Linoff [10] a modo de complemento.

Los procesos que se pueden llevar a cabo con las actuales herramientas computacionales disponibles en estaciones de trabajo personales, permiten procesar grandes cantidades de datos a velocidades de unos cuantos Gigaflops, revelando relaciones entre las variables que antes era imposible realizar. La

[9] http://www.crisp-dm.org/CRISPWP-0800.pdf y http://www.cs.ualberta.ca/~yli/CRISPDM.ppt

[10] M. Berry, G. Linoff, "Data mining techniques for marketing, sales and customer relationship management", Wiley Publishing, Inc. Indianapolis, 2004, pp. 43 – 86. [11]

minería de datos por lo tanto soportada en estas tecnologías difiere de los análisis estadísticos en que esta última involucra análisis exploratorio de datos (*Exploratory Data Análisis, EDA*) y no es orientada por hipótesis preliminares[11].

Schumann,[11] en su publicación prueba que CRISP-DM, como metodología usada en el mundo de los negocios, puede ser transferida para su uso en instituciones educativas produciendo información operacional útil para guiar a los estudiantes en los logros planteados por sus instructores. Los Programas computacionales que contengan algoritmos de clasificación y árbol de regresión pueden con gran certeza predecir el comportamiento de una cohorte de estudiantes, incluso si el modelo elaborado proviene de otra cohorte de estudiantes. Usando estos análisis los directivos de una universidad pueden identificar variables importantes que influyen en los resultados de los estudiantes.

La razón citada en el párrafo anterior y las siguientes razones fueron base para la elección de estas metodologías:

- Existe gran documentación y aceptación en el mercado y como el nombre CRIPS lo indica, se considera un estándar metodológico para llevar a cabo proyectos de minería de datos.

- La metodología planteada por Berry – Linoff coincide en gran parte con la metodología CRISP-DM, difiriendo en el orden como se realizan cada uno de los pasos, pero completándola en otros casos, tal cómo se explica en los párrafos siguientes.

- Es relativamente fácil la implementación de un proyecto utilizando estas metodologías, dada su simplicidad.

- La mayoría de herramientas de minería de datos (como por ejemplo Oracle Data Mining, Mine Set, Clementine, etc) permiten llevar a cabo el proceso según las indicaciones definidas por estas metodologías.

CRISP-DM. Ésta metodología se describe de forma jerárquica consistente en un conjunto de tareas en cuatro niveles de abstracción (desde lo general a lo específico): fase, tarea genérica, tarea especializada e instancia de proceso. Más en detalle ésta metodología en su cido de vida se describe de la siguiente manera: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación que se describen brevemente a continuación, tal como se muestra en la figura 6.

[11] Schumann, Jeffrey A. PhD. Data mining methodologies in educational organizations. UNIVERSITY OF CONNECTICUT. 2005. [18]

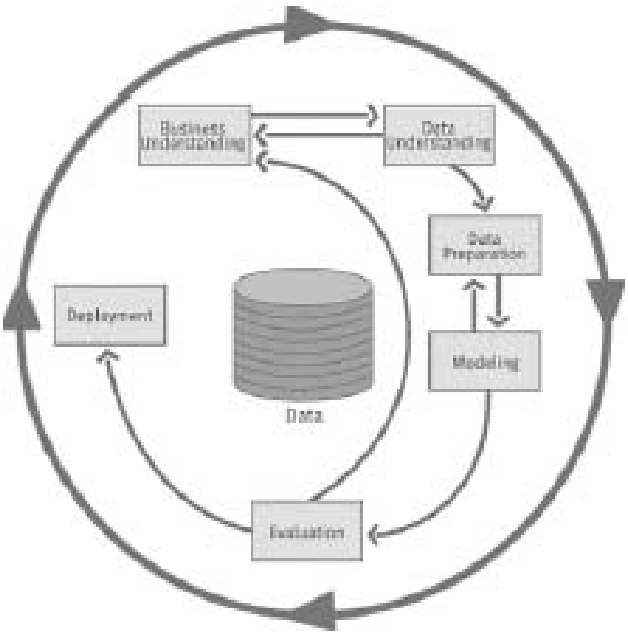

Figura 6. Metodología CRISP-DM [17]

6.1 Entendimiento del Negocio

Se enfoca en entender los objetivos del proyecto desde una perspectiva de negocios, convirtiendo este conocimiento en una definición del problema de minería de datos y un plan para alcanzarlo. Para poder entender cuáles datos deberán luego ser analizados, y cómo, es vital para los analistas el completo entendimiento del negocio para el cual están buscando soluciones.

Incluye las siguientes tareas:

1. *Determinar los objetivos del negocio:* El primer objetivo del analista es entender, desde una perspectiva de negocios, lo que el usuario realmente desea lograr, balanceando objetivos contrapuestos que pueden existir. Una posible consecuencia de descuidar este paso es gastar mucho esfuerzo en producir la respuesta correcta a problemas equivocados.

2. *Evaluar la situación:* Implica la definición de recursos, restricciones y otros factores que deben ser considerados para determinar los objetivos del análisis de datos y el plan del proyecto.

3. *Determinar los objetivos de la minería de datos:* Un objetivo de negocio expresa los objetivos en terminología de negocios. Un objetivo de minería de datos expresa los objetivos en términos técnicos. Por ejemplo, el objetivo de negocios puede ser “Incrementar la cantidad de estudiantes existentes en un período lectivo dado”. Un objetivo de minería de datos seria “Predecir cuantas

asignaturas electivas un estudiante cursará, dada la cantidad de asignaturas tomadas en los últimos tres semestres, información demográfica (edad, salario, ciudad, etc.) y el valor del semestre."

El complemento en esta fase lo da Berry- Linoff[12] cuando plantean que este proceso debe convertirse en una o unas de las siguientes tareas:

- Clasificación, estimación, predicción, minería de datos directa (aprendizaje supervisado) debe haber siempre una variable objetivo. Algo a ser estimado, clasificado o predecido. Para construir un clasificador se comienza con un grupo predefinido de clases y ejemplos que ya han sido correctamente clasificados. Para construir un estimador, se comienza con datos históricos donde los valores de la variable objetivo ya se conocen. El modelamiento consiste en encontrar las reglas que explican los valores conocidos de la variable objetivo.

- Grupo de afinidad y conglomerados minería de datos indirecta (aprendizaje no supervisado). No hay variable objetivo. La tarea de minería de datos es encontrar todos los patrones que no están atados a alguna variable. En la funcionalidad de conglomerados (*clustering*) encuentra grupos de registros similares sin instrucción acerca de cuáles variables deben ser consideradas como las más importantes. Es descriptivo por naturaleza.

- Descripción y tipificación (*profiling*). Tanto directo como indirecto.

4. *Producir un plan de proyecto:* Describir un plan de proyecto para alcanzar los objetivos de minería de datos y por ende los del negocio. El plan debe especificar el conjunto de pasos a ser realizados durante el resto del proyecto incluyendo una evaluación inicial de las herramientas y técnicas.

5. *Definir la población:*

   Establecer a que o quienes son los objetos del proceso a realizar. Por ejemplo en el presente trabajo la población objetivo son los estudiantes de la universidad (toda la población), tanto egresados como no egresados. Ésta tarea permite segmentar o definir más precisamente el objeto de estudio filtrando aquellos elementos que no ayudan mucho en el estudio o seleccionando inicialmente a quienes se va aplicar.

6. *¿Cómo se muestran los resultados?*
   Cuando el objetivo es obtener una identificación significativa los entregables pueden ser graficas o diagramas. Cuando es una prueba de conceptos o proyecto piloto el entregable puede ser una lista de estudiantes que recibirán diferente tratamiento en el sistema. Cuando el proyecto está actualmente en curso, entonces el proyecto es parte de un esfuerzo de administración de la relación analítica con el usuario, el entregable es un programa o grupos de

[12] Ibid,, Berry-linoff.

programas que pueden ejecutarse en una base regular para medir un subgrupo de población de usuarios con software adicional para administrar los modelos y valores en el tiempo.

6.2 Entendimiento de los datos.

Comienza con una recolección inicial de los datos. El analista procede a incrementar la familiaridad con ellos, a identificar problemas de calidad, a descubrir conocimiento inicial, o a detectar subconjuntos interesantes para luego formar hipótesis acerca de información oculta.

Involucra los siguientes cuatro pasos:

1. *Recolección inicial de datos*: Abarca la adquisición o el acceso a los datos. Esta recolección inicial incluye la carga de datos si es necesario el entendimiento del negocio. Por ejemplo, si se aplica una herramienta especifica para entendimiento del negocio, se debe cargar los datos en dicha herramienta. En esta tarea [Berry-Linoff] plantea que es importante seleccionar los datos apropiados, como base para todo el proceso. Es por esa la razón de tener un conocimiento definido acerca de que se quiere hacer con minería de datos.

2. *Describir los datos:* Examinar las propiedades "superficiales" de los datos adquiridos y reportar los resultados.

3. *Explorar los datos:* Implica la consulta de los datos, la visualización y los reportes. Por ejemplo, el analista de datos puede consultar los datos para descubrir las carreras que los estudiantes de un determinado grupo social usualmente cursan. O el analista puede ejecutar análisis de visualización para descubrir patrones de estudiantes desertores. Luego se elaboran los reportes que describen los resultados obtenidos, o las hipótesis iniciales y los potenciales impactos para el resto del proyecto. La tarea se puede llevar a cabo al examinar distribuciones, comparar valores con descripciones y es muy importante "formular muchas preguntas"

4. *Verificar la calidad de los datos:* Examinar la calidad de los datos respondiendo preguntas como ¿están todos los datos necesarios? ¿son correctos? ¿hay errores en los mismos? etc.

6.3 Preparación de los datos

Cubre todas las actividades para la construcción del conjunto de datos final (dataset), es decir, los datos que alimentarán la herramienta de modelado. Las

tareas incluyen selección de tablas, registros y atributos y también la transformación y limpieza de los datos a utilizar. Los pasos son:

1. *Selección de datos:* Decidir cuáles datos se usarán para el análisis. Los criterios incluyen la relevancia con los objetivos de minería de datos, calidad, restricciones técnicas, etc. La selección de los datos cubre la selección de atributos y la selección de registros en una tabla.

2. *Limpieza de datos*: Llevar la calidad de los datos al nivel requerido por las técnicas seleccionadas. Esto puede incluir, por ejemplo, la estimación de valores faltantes, variables categóricas con muchos valores, variables numéricas con distribuciones *Skewed* o *outliers*, valores omitidos, valores que significan cosas diferentes en el tiempo, codificación de datos inconsistente.

3. *Construcción de datos:* Esta tarea incluye las operaciones constructivas de datos como la generación de atributos derivados, inserción de nuevos registros o transformación de valores de atributos existentes.

4. *Integración de datos:* Implica la combinación de información proveniente de múltiples tablas o registros para crear nuevos registros o valores.

5. *Formateo de datos:* En algunos casos se puede necesitar cambiar el formato o diseño de los datos. Estos cambios pueden ser simples (reducir una cadena de texto a una longitud máxima) o complejos (reorganizar la información). A veces estos cambios se necesitan para poder adaptar los datos a las herramientas de minería de datos, capturar tendencias, crear prorrateo o combinaciones de variables, convertir conteos en proporciones.

6.4 Modelado

Se seleccionan y aplican varias técnicas y algoritmos de modelado, sus parámetros son ajustados para lograr valores óptimos. En general, se pueden aplicar varios tipos de algoritmos al mismo problema de minería de datos. Algunos de ellos tienen requisitos en cuanto a la forma de los datos. Así, puede ser necesario volver a la fase de preparación de los datos.

Los pasos de modelación incluyen:

1. *Seleccionar la técnica y algoritmo de modelado:* Seleccionar la técnica y algoritmo a usar. Por ejemplo, un árbol de decisión o una red neuronal con propagación hacia atrás. Si se van a aplicar múltiples técnicas, se debe realizar esta tarea para cada una de las técnicas por separado.

2. *Generar diseño de prueba:* Antes de construir el modelo, necesitamos generar un procedimiento para probar la calidad y validez del modelo. Por ejemplo, en una clasificación es común utilizar las tasas de error como medidas de calidad

de los modelos de minería de datos. Así, en general, se separa el grupo de datos (*dataset*), en un conjunto de entrenamiento y otro de prueba, se construye el modelo sobre el de entrenamiento y se estima la calidad sobre el de prueba.

3. *Construir modelo:* Ejecutar la herramienta de minería de datos sobre el grupo de datos preparado para crear uno o más modelos.

4. *Evaluar modelo:* Se interpretan los modelos según el conocimiento del dominio, el criterio de éxito y el diseño de prueba del mismo. El ingeniero de minería de datos juzga el éxito del modelado más técnicamente, se contacta con los expertos en el negocio para analizar los resultados desde un contexto de negocios. Esta tarea sólo considera modelos, mientras que la evaluación toma en cuenta todos los otros resultados que se produjeron durante el proyecto.

6.5 Evaluación

Se evalúa el modelo y se revisa la construcción del mismo para estar seguros de que el mismo alcanza los objetivos del negocio. Es crítico determinar si algún aspecto del negocio no fue lo suficientemente considerado.

Al finalizar, el líder del proyecto debe decidir cómo usar los resultados de minería de datos. Los pasos claves son:

1. *Evaluación de resultados:* Este paso evalúa el grado con el cual el modelo satisface los objetivos de negocios y busca determinar si existe alguna razón del negocio por la cual este modelo podría ser deficiente.

2. *Revisión del proceso:* Se revisa el proceso de elaboración del modelo para determinar si existe algún factor o tarea que ha sido pasada por alto. También para analizar si se construyó correctamente el modelo, etc.

3. *Determinar próximos pasos:* El líder del proyecto debe decidir si finaliza el proyecto y prosigue con la implementación del mismo o si inicia nuevas iteraciones o nuevos proyectos de minería de datos.

4. *Revisión del proyecto:* Se evalúa qué se hizo bien, qué se hizo mal y qué necesita mejorarse.

El aspecto menos favorable de la metodología radica en considerar los paquetes de software y técnicas en la misma fase (entendimiento del negocio). Si las herramientas y las técnicas son combinadas y seleccionadas simultáneamente, las técnicas pueden elegirse porque son soportadas por las herramientas y no porque son las más relevantes para el propósito del estudio o porque son necesarias. Esto puede ocasionar que los objetivos de la organización sean parcialmente considerados. Las técnicas bajo CRISP-DM pueden ser aplicadas porque se

encuentran incorporadas en las herramientas disponibles para la organización y no porque realmente sean necesarias. De esta manera, los resultados obtenidos pueden no corresponderse correctamente con los principales objetivos de la organización y los modelos generados puede que no representen verdaderamente el comportamiento de entidades para las cuales el estudio fue pretendido en primer lugar [13][6].

[13] Solarte, José. A Proposed Data Mining Methodology and its Application to Industrial Engineering. Agosto 2002. [8]

7 HERRAMIENTAS DISPONIBLES PARA LA SOLUCION DEL PROBLEMA

7.1 Sistemas Operacionales y Administradores de bases de datos (DBMS).

Para la realización del desarrollo de una herramienta informática que dé el soporte necesario se tiene el soporte de los diferentes sistemas operacionales comerciales y de amplia aceptación en el mercado como son: Microsoft Windows, Linux en diferentes versiones (Red Hat, Mandrake, SUSE, ubuntu, etc), Sun Solaris, y SCO UNIX. Estos sistemas soportan a su vez varios administradores de bases de datos tales como ORACLE, DB2, MS SqlServer, NCR Teradata, Sybase que tienen facilidades tanto de bases de datos, como de bodegas de datos y minería de datos.

Por restricciones de tiempo y para algunos casos de dinero, no se pudo realizar una evaluación exhaustiva de cada una de las herramientas citadas anteriormente, aunque sí se miró la funcionalidad de herramientas como NCR Teradata, SPSS Clementine y Oracle, en cuanto a sistemas robustos y Purple Mine set y WEKA como sistemas de uso libre. Se escogió hacer la siguiente descripción relacionada con Oracle debido a que es una herramienta potente, bastante probada y como una restricción de implantación del prototipo ya que la base de datos sobre la cual se aplicara y desarrollara está en ese DBMS.

7.2 Oracle data Mining[14] (ODM).

Esta integrada a la funcionalidad del motor de la base de datos relacional de Oracle. Los algoritmos operan en forma nativa en dicho ambiente, eliminando la necesidad de extraer y transferir los datos. Los procesos relacionados con la minería de datos pueden correr de manera asincrónica e independiente de cualquier interface de usuario, y por lo tanto los analistas de datos, pueden construir modelos y metodologías y aplicarlos a las aplicaciones desarrolladas por ellos y ser puestas rápidamente en producción. El ODM puede ejecutarse en acceso multi-sesión de usuario único utilizando interface JAVA. Cuando se utiliza el PL/SQL el proceso se hace sincrónicamente.

La minería de datos Oracle permite la ejecución de las siguientes funciones:
En minería de datos supervisada la clasificación, regresión, puntuación de atributos y detección de anomalías.

En minería de datos no supervisada la clasificación por conglomerados (*clustering*), modelos de asociación (ej. El análisis de la canasta de mercado,

[14] Conceptos de Oracle data mining. Manual 10*g* Release 2 (10.2) referencia: **B14339-01 [**19**]**

Market Basket analysis) y extracción de nuevas características (*feature extraction*) como resultado de la combinación de otros atributos.

Además la minería de datos Oracle permite hacer minería sobre una o más columnas de texto y soporta algoritmos especializados de búsqueda secuencial y alineamiento (*BLAST Algorithms*) usados para detectar similaridades entre un nucleótido y secuencias de amino-ácidos.

La versión de Oracle 10g R2, incluye los algoritmos de árboles de decisión y soporte de máquina vectorial de una clase (*One-Class Support Vector Machine algorithm*) junto con el aprendizaje activo. Incluye además la herramienta *Oracle data miner*, la cual es una interface gráfica de usuario final y otras características para el lenguaje PL/SQL.

7.2.1 ALGORITMOS DE MINERÍA DE DATOS UTILIZADOS EN ODM.[15]

Como ya se mencionó, la minería de datos ODM, hace uso de la minería de datos tanto supervisada como la no supervisada. Para la minería de datos supervisada (que necesita una variable objetivo) usa para los algoritmos de clasificación: árboles de decisión, Naive Bayes y SVM; algoritmos de regresión y detección de anomalías. En cuanto a la minería de datos no supervisada para realizar conglomerados (*clustering*) utiliza *k-means y scoring*; asociación y extracción de características de los cuales se hace una breve descripción a continuación.

7.2.1.1 Minería Supervisada:

Naive Bayes[16]**:** es una técnica de clasificación y predicción que construye modelos determinando la probabilidad de posibles resultados. Utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones. Este algoritmo halla resultados binarios o multiclase. En los problemas binarios, cada registro cumplirá o no el comportamiento modelado. Por ejemplo, ¿Es probable que el estudiante termine sus estudios?
Respuesta: Si, con un 45% de probabilidad.

Naive Bayes puede hacer predicciones para problemas multiclase, en los cuales hay varios resultados posibles. Por ejemplo, ¿En cuál de los tipos de colegios de secundaria (publico, privado, religioso. técnico) encaja mejor este estudiante?
Respuesta: Religioso, con el 25% de probabilidad.

---

[15] Oracle® Data Mining Concepts 10*g* Release 2 (10.2) B14339-01. Junio de 2005.

[16] MontakeFlyerForumB. PDF Oracle Data Mining. Publicado el 17/2/04 [21]

Las predicciones binarias y multiclase conjuntamente cubren un gran rango de problemas de negocio, incluyendo respuesta a campañas, ofertas de ventas, detección de fraude, predicción de rentabilidad, perfilado de clientes, etc.

Una vez creados los modelos Naive Bayes, a cada registro de datos se le puede dar un puntaje. La puntuación es el proceso de predicción de resultados, y puede hacerse en modo batch o bajo demanda. En modo batch el algoritmo recorre una tabla y va almacenando las predicciones en otra tabla; bajo demanda el algoritmo da puntaje a un sólo registro y devuelve la predicción, que puede utilizarse directamente en la aplicación que haya pedido esta puntuación.

El algoritmo esta basado en el teorema de Bayes, el cual deriva la probabilidad de predicción de una evidencia subyacente. Se basa en que la probabilidad de un evento A ocurre dado que un evento B ha ocurrido, es proporcional a la probabilidad que el evento B ocurra dado que el evento A ha ocurrido multiplicado por la probabilidad de ocurrencia del evento A

$$P(a/b) = ((P(b/a) * P(b)$$

Naive Bayes asume que cada atributo es condicionalmente independiente de otros, es decir que dado un valor particular, la distribución de cada predictor es independiente de los otros predictores. En la práctica, asumir esta situación de independencia, aún cuando sea violada no degrada la precisión de la predicción en forma significativa.

ÁRBOLES DE DECISIÓN. Esta técnica se utiliza para encontrar las razones por la cuales los datos se clasifican en determinados grupos. Por ejemplo, determinar las razones por las que los alumnos tienen buenos o malos rendimientos.

Los registros de entrada se componen de un conjunto de valores para distintos factores y de una clasificación. Si se analizan los valores de todos los registros se pueden determinar comportamientos comunes que contribuyen a realizar la clasificación. El modelo de clasificación resultante se puede utilizar para predecir las clases de registros que no han sido clasificados.

Las reglas de los árboles de decisión proveen transparencia al modelo, de tal manera que quien trabaje con él (usuario, analista de Mercado, analista del negocio, etc) pueda entender las bases para la predicción y por lo tanto poderlas explicar cuando sea necesario.

Los árboles de decisión construyen modelos para predicciones binarias y multiclase produciendo modelos precisos y fácilmente interpretables con relativa poca intervención del usuario. Se implementan de tal forma que se pueden manejar datos en tablas, que tienen criterios de particionamiento y terminación con manejo automático de valores nulos u omitidos.

Esta técnica utiliza un algoritmo de inducción en árbol, describiendo la distribución subyacente de los datos. Realiza un ajuste proporcional con respecto al número

de datos de ejemplo y al número de atributos que se encuentran en la fuente de los datos. Los parámetros que se deben proporcionar para realizar la clasificación son los siguientes.

1. Datos de entrada. Objeto (tabla) de datos de entrada.
2. Filtrado de registros
3. Parámetros de modalidad
   - Máximos niveles del árbol: limita el número de niveles de nodos del árbol de decisión.
   - Pureza máxima por nodo interno: detiene ulteriores divisiones del árbol cuando se ha alcanzado cierto nivel de pureza.
   - Mínimo de registros por nodo interno: determina el número de registros que se deben incluir en un nodo antes de pasar al siguiente.
4. Optimización de la función de Minería de Datos. Optimización en tiempo o espacio de almacenamiento.
5. Campos de entrada. Factores que se tendrán en cuenta para la clasificación.
6. Pesos de campos. Se utiliza para ponderar diferentes factores en la clasificación.
7. Matriz de errores. Permite realizar una ponderación de errores asimétrica asignando más o menos pesos a ciertas descalificaciones. Por ejemplo: en el caso de clasificar el rendimiento en alto, medio, bajo. Un error que clasifique un alumno como de alto rendimiento cuando en realidad debería clasificarlo como de bajo rendimiento debe estar más penalizado que si el error de clasificación fuera de rendimiento medio a bajo.
8. Datos de salida. Objeto (tabla) de datos de salida donde se almacenan los resultados.

7.2.1.2 Minería no supervisada.

Conglomenrados. (Clustering)

Se usan para encontrar agrupamientos naturales que están ocultos en los datos. Los grupos son colecciones de datos que son similares en algún sentido a otros. Un buen método de agrupamiento produce grupos de alta calidad que aseguran que la similaridad entre grupos es baja y que la similaridad de los elementos del grupo es alta. En otras palabras los elementos de un grupo se parecen mucho entre ellos y son muy diferentes a los de otro grupo.

El agrupamiento puede ser un paso útil en el pre-procesamiento de datos para identificar grupos homogéneos sobre los cuales construir modelos supervisados. Los modelos no supervisados son diferentes de los modelos supervisados en que el resultado del proceso no esta guiado por un resultado conocido, es decir que no hay una variable objetivo. El promedio de error entre este y el valor predicho se puede calcular en la construcción del modelo. El modelo resultante puede ser usado para asignar identificadores de grupo a puntos de datos. En minería de datos Oracle (*ODM*) un conglomerado está caracterizado por un centroide, histogramas de atributos y la ubicación del cluster en la jerarquía del árbol. ODM

genera agrupamientos jerárquicos usando una versión mejorada del algoritmo *K-means* y el algoritmo de agrupamiento por particionamiento ortogonal () (orthogonal partitioning clustering algorithm - O-cluster) propiedad de Oracle. Los grupos detectados por estos algoritmos se usan para crear reglas que capturan las principales características de los datos asignados a cada grupo. Por ejemplo un grupo de datos asignado a un grupo determinado X, puede tener las siguientes reglas: 18 meses <tiempo_receso > 6 meses y semestre cursado <= 3 entonces grupo X.

El agrupamiento también sirve para generar modelos de probabilidad Bayesiana los cuales son usados para la valoración de asignación de los puntos de datos asignados a los grupos.

Algoritmos para conglomerados. En particular ODM tiene los siguientes algoritmos: k-means mejorado y agrupamiento por particionamiento ortogonal (O-cluster).

K-means mejorado**.** El algoritmo K-means es un algoritmo que se basa en medidas de distancia para encontrar la similaridad de los puntos de datos y particiona los datos en un determinado número de grupos, teniendo presente que haya suficientes casos distintos. La medida puede ser Euclidiana, Coseno, o por medio de la distancia rápida del coseno. Los puntos se asignan al grupo más cercano de acuerdo a la medida de distancia usada. En el caso de ODM, el algoritmo k-means tiene las siguientes características:

Construye los modelos de una manera jerárquica, de arriba abajo, usando divisiones binarias y refinando todos los nodos al final (similar al algoritmo de bisección K-means). El centroide de los nodos más internos en la jerarquía se actualiza para reflejar los cambios a medida que el árbol evoluciona.

Basado en un valor dado por el usuario en una interface, el algoritmo crece de un nodo a la vez (acercamiento no balanceado). Eligiendo el nodo con la mayor varianza, se divide para incrementar el tamaño del árbol hasta alcanzar el número de grupos deseado.

El algoritmo también provee la valoración probabilística y la asignación de datos a los grupos y devuelve para cada centroide un histograma (para cada atributo) y una regla que describe el super-marco que envuelve la mayoría de los datos asignados a cada grupo. El centroide es la moda para los atributos categóricos de la media y la varianza de atributos numéricos. Esta aproximación al algoritmo k-means, evita la necesidad de construir muchos modelos y provee resultados de conglomerados superiores en consistencia al algoritmo k-means tradicional.

La implementación en ODM del algoritmo soporta tanto datos categóricos, como datos numéricos. En el caso de los datos numéricos se recomienda la normalización de los datos, aunque sí los datos son escasos o se han omitido al azar, se podrá realizar alguna asignación a los datos por medio del tratamiento de

valores omitidos y remplazar los valores no nulos por valores nulos. Una forma sencilla de tratar los valores omitidos es usar la media para valores numéricos y la moda para valores categóricos. En el caso que no realice este proceso el algoritmo tratará los valores en forma incorrecta.

Puntuación (*scoring*). Los grupos encontrados por el algoritmo mejorado de k-means, se usan para generar el modelo de probabilidad Bayesiano, que es el usado para la valorar cada punto de dato del grupo. Adicionalmente este algoritmo se puede interpretar como un modelo en el que los datos resultantes son componentes de una distribución normal multivariada esférica (*spherical multivariate normal distributions*) con la misma varianza para todos los componentes.

Asociación. Los modelos de asociación se usan frecuentemente para el análisis de la cesta del mercado, el cual intenta descubrir la correlación en un grupo de datos. Este análisis se usa para el diseño de catálogos en mercadeo, mercadeo directo y otros procesos de decisión relacionados con los negocios. Una regla típica de asociación infiere por ejemplo que "80 % de los estudiantes que pierden física y álgebra también perderán estadística".

Los modelos de asociación capturan la co-ocurrencia de ítems o eventos en grandes cantidades de datos de las transacciones de los clientes. Esta labor se facilita con la disponibilidad de los códigos de barras, al recoger el detalle y almacenar grandes cantidades de datos de ventas. Los modelos de asociación se diseñaron inicialmente para estos procesos aunque tienen muchas otras aplicaciones. Encontrar las reglas de asociación tiene mucho valor para las promociones de mercadeo relacionadas con órdenes por correo y mercadeo cruzado (cross-marketing) y también en otras aplicaciones como diseño de catálogos, ventas agregadas, segmentación de clientes, personalización de páginas Web y mercadeo objetivo.

Tradicionalmente los modelos de asociación se han usado para:

- Descubrir tendencias en las transacciones de los clientes;
- Encontrar todas las combinaciones de ítems, es mayor que el soporte mínimo especificado;
- Para predecir los accesos a las páginas Web en cuanto a la personalización de dichos accesos;

Extracción de características. La extracción de características crea un grupo de características basadas en los datos originales. Una característica es una combinación de atributos que son de especial interés, capturando detalles importantes de los datos.

Algunas aplicaciones de la extracción de características son el análisis, semántica latente, comprensión de datos, descomposición y proyección de datos y reconocimiento de patrones. También se puede usar para mejorar la velocidad y

efectividad del aprendizaje supervisado. Por ejemplo, se puede usar para extraer temas de una colección de documentos, donde estos se representan por un grupo de palabras clave, con sus frecuencias de ocurrencias respectivas. Cada tema (característica) se representa por una combinación de palabras clave, por lo tanto el documento en una colección, se puede expresar en términos de temas descubiertos.

Algoritmos para Extracción de características.

ODM usa el algoritmo de la matriz n-factorización no negativa para la extracción de características[17] (NMF). Este algoritmo descompone los datos multivariados, creando un número de características que terminan en una representación reducida de los datos originales. El NMF descompone una matriz de datos V en el producto de dos matrices de bajo rango W y H, así que V es aproximadamente W veces H. El algoritmo usa un proceso iterativo para modificar el valor inicial de W y H, para que el producto se aproxime a V. El proceso termina cuando el error de aproximación converge o cuando se alcanza un determinado número de iteraciones. Cada característica es una combinación lineal de los atributos originales del grupo; los coeficientes de esta combinación lineal son no negativos.

Algoritmo de soporte de la maquina vectorial (Support Vector Machine Algorithm, SVM). El estado del arte en los algoritmos de clasificación y regresión, tiene muchas propiedades de regularización, al maximizar la precisión de la predicción, mientras que automáticamente evita el sobre-entrenamiento de un grupo de datos en entrenamiento, propiedades que están presentes actualmente en los modelos SVM. En su funcionamiento general, éste algoritmo envía los datos de entrada al espacio del kernel donde construye un modelo lineal.

Existen principalmente dos algoritmos en SVM: de clasificación y de regresión. Un modelo SVM de clasificación intenta separar las clases objetivo con el más amplio margen posible. Un modelo SVM de regresión trata de encontrar una función continua tal que maximice el número puntos de datos que están a una distancia epsilon. La escogencia de diferentes tipos de kernels y parámetros puede generar diferentes fronteras de decisión (clasificación) o funciones de aproximación (regresión).

Particularmente el ODM SVM, soporta dos tipos de kernel: lineal y Gausiano. El SVM se comporta bien con aplicaciones del mundo real tal como clasificación de textos, reconocimiento de caracteres escritos, clasificación de imágenes, análisis de bio-informática y bio-secuencia. No hay un límite de atributos y cardinalidad objetivo, solamente esta aquella impuesta por el hardware. El SVM es el algoritmo preferido para el análisis de datos escasos (*sparce*).

[17] D. D. Lee y H. S. Seung, Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* (401, pages 788-791, 1999. [22]

Muestreo para la clasificación. El SVM automáticamente ejecuta el muestreo estratificado durante la construcción del modelo. Explora la totalidad de los datos construidos y elige la muestra más balanceada entre los valores objetivo.

Selección automática del núcleo (kernel). SVM automáticamente determina el tipo de núcleo adecuado basado en las características de los datos. Esta selección puede ser modificada especificando explícitamente el tipo de núcleo.

7.3 Análisis de retención (curvas de sobrevivencia, y curvas de retención) y análisis de amenazas.

Otra técnica recientemente empleada en mercadeo utilizando herramientas de minería de datos, muy relacionada con la retención de clientes, es el Análisis de supervivencia o análisis de eventos en el tiempo (time to event analysis). Esta está orientada a entender a los clientes cuando hay motivos por los cuales preocuparse (como es el caso de la deserción aquí estudiada), relacionados estos con la ocupación que están haciendo de los servicios de la organización y determinar cuáles factores de los que tienen que ver con una relación comercial tienen el mayor efecto en la permanencia de un cliente.

La característica de esta última técnica es la de proveer el entendimiento de eventos en el tiempo tales como [11]:

- Cuándo será que un estudiante se retirará.
- Cuál será la siguiente ocasión que un estudiante se cambiara a otra carrera o a otra universidad.
- Cuál la siguiente ocasión que un estudiante realizará un nuevo curso de posgrado en la universidad.
- Cuáles son los factores en la relación comercial que incrementan o disminuyen el uso de los recursos por parte de los estudiantes.

Hay dos valores muy importantes que soportan la información de retención o sobrevivencia de los estudiantes: la fecha de inicio y la fecha de finalización de la carrera. En algunas aplicaciones estos son bastante evidentes, como por ejemplo el inicio y finalización de un semestre académico, un curso de capacitación o inscripción a alguna revista o periódico. Pero en otras aplicaciones, como las transaccionales, estos datos no son muy evidentes. Con los datos proporcionados por los periodos de retención de clientes, se puede hacer una grafica o curva de retención, la cual mostrara el porcentaje de estudiantes que se retienen para un periodo de tiempo particular (ver figura 7). En esta curva los estudiantes que están en los mayores períodos de retención también lo están en los menores, es decir tiene un comportamiento de Histograma y permite establecer cuándo la mitad de los estudiantes sale o deja de tener vínculos con la universidad (10 semestres, en la gráfica) y también a partir de ella, el tiempo promedio de retención de un

estudiante (ya que el área bajo la curva es la cantidad de estudiantes dividido por la cantidad total de estudiantes (aproximadamente 5 semestres en la grafica).

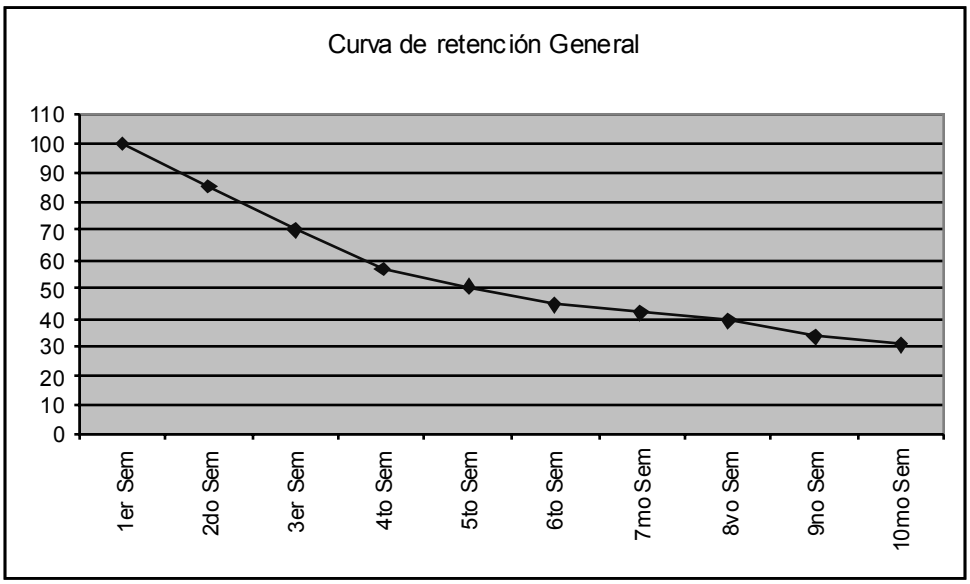

Figura 7. Curva de retención de estudiantes por semestre con duración total de 10 semestres.

Un análisis de retención se complementa en forma más precisa con el análisis de amenazas (Hazard analysis) el cual posteriormente podrá servir de base para el análisis de sobrevivencia. Mientras que el primero permite responder a preguntas como ¿cuántos estudiantes que han sobrevivido a una cantidad de tiempo tx, saldrán o se retiraran en el tiempo tx +1? la segunda determina la probabilidad que hay que un estudiante permanezca después de ese periodo de tiempo (los valores de sobrevivencia normalmente se calculan a partir de los valores dados por el análisis de amenazas).

8 ESTADO DEL ARTE: SISTEMAS PARA RETENCION Y DESERCION UNIVERSITARIA

El problema de deserción es un fenómeno que ha sido tratado en el "Observatorio de la Educación Superior en América Latina y el Caribe" del Instituto Internacional para la Educación Superior en América Latina y el Caribe (IESALC / UNESCO), debido a su complejidad, su magnitud e implicación social y a través del cual se han desarrollado diversos estudios en toda América y en Europa, con el fin de caracterizarlo, tratarlo y plantearle soluciones.

En enero de 2004 SPSS, compañía dedicada a elaborar software para análisis estadístico y CLEMENTINE, software para minería de datos, publicó a través de JING LUAN, PHD, Jefe de planeación y registro, el artículo "Data mining application in Higher education"[1] en el cual se plantean los beneficios del uso de la minería de datos, en entidades universitarias. En él se hace una comparación de los procesos llevados a cabo en entidades de carácter comercial que también utilizan herramientas de minería de datos. Luego de presentar tres casos de estudio, el autor concluye "estas herramientas le permiten a una universidad un mejor uso y planeación de los recursos y la predicción de algunos comportamientos tales como procesos de aprendizaje, transferencia entre carreras y relaciones de mercadeo con los estudiantes" [1].

En el IV Coloquio Internacional sobre Gestión Universitaria en América del Sur ALIANZAS ESTRATÉGICAS, INTEGRACIÓN Y GESTIÓN UNIVERSITARIA 8-10 de diciembre de 2004, Florianópolis, Santa Catarina – Brasil, el Ing. Diego D. Gregoraz y el Dr. Julio C. Durand, presentaron la ponencia "Explotación de la información académica para la mejora continua", donde entre otros aspectos orientados a mejorar los procesos administrativos de la universidad, incluyeron la minería de datos como herramienta "que está en la intersección de varias disciplinas como estadística, aprendizaje automático (*machine learning*) y manejo de bases de datos, entre otras". A través de un ejemplo mostraron cómo se pudo establecer la correlación de diferentes variables (por ejemplo ubicación geográfica y resultados en diferentes asignaturas) para encontrar comportamientos ocultos de la deserción de los estudiantes.

En la Argentina, Universidad Nacional de la Patagonia, UNPA, un grupo de docentes realizó un estudio debido a la preocupación existente ante la deserción de estudiantes. "Inicialmente se realizó un estudio cuantitativo que en una primera etapa estableció el perfil del alumno que ingresaba y si esto tenía relación con el abandono. La segunda etapa fue de comprensión de las razones y en esta instancia, se fue directamente a la búsqueda de los alumnos desertores para que explicaran los motivos. Otra instancia se fundó en la percepción de esos motivos, donde descubrieron por ejemplo, entre otros, el aspecto metodológico de estudio, la relación con los profesores, las exigencias y falencias de aprendizaje" [2].

En la universidad de Río, provincia de Córdoba, Argentina, también se ha llevado a cabo un proceso similar de tipo investigativo- descriptivo, teniendo en cuenta que el proceso de deserción universitaria comprende tres términos “proceso de selección, medida del rendimiento académico, y eficacia del sistema educativo" [3]. El primero se enmarca en el enfoque sociológico, según el cual la selección que se opera en la Enseñanza Superior, se constituye en un filtro social que frena la movilidad. El segundo, en la Universidad se debiera abordar tres dimensiones: éxito, retraso y abandono. Y el tercero, la deserción sólo da cuenta que en algunos estudios se registra una mayor tendencia al abandono en las instituciones que no tienen examen de ingreso [3]. En Uruguay [4], y en Europa [5] las preocupaciones son similares y los estudios realizados pretenden explicar y encontrar soluciones al tema.

En México (2001), por ejemplo, se ha realizado un estudio sobre los recién egresados de la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II, para observar si sus recién titulados se insertaban en actividades profesionales relacionadas con sus estudios y, en caso negativo, se buscaba saber el perfil que caracterizó a los ex alumnos durante su estancia en la universidad.

El objetivo era saber si con los planes de estudio de la universidad y el aprovechamiento del alumno se hacía una buena inserción laboral o si existían otras variables que participaban en el proceso.

Dentro de la información considerada estaba:
- el sexo
- la edad
- la escuela de procedencia
- el desempeño académico
- la zona económica donde tenía su vivienda
- la actividad profesional, entre otras variables.

Mediante la aplicación de conjuntos aproximados se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral, que son citadas de acuerdo con su importancia:
- zona económica donde habitaba el estudiante
- colegio de donde provenía
- nota al ingresar
- promedio final al salir de la carrera.

A partir de estos resultados, la universidad tuvo que hacer un estudio socioeconómico sobre grupos de alumnos que pertenecían a las clases económicas bajas para dar posibles soluciones, debido a que tres de las cuatro variables no dependían de la universidad [2].

La tecnología de minería de datos, brinda una visión de los problemas innovadora. En el ejemplo anterior, el objetivo es analizar la problemática de la inserción

laboral de los egresados considerando variables de diverso tipo.

Posiblemente, si no se consideraran las posibilidades que brinda la minería d e datos, sería muy difícil obtener información útil a partir de la importante cantidad de datos que se conservan de los alumnos egresados.

Para complementar la lista de posibles problemas en donde es importante la obtención de conocimiento, se podría considerar analizar la deserción paulatina de estudiantes que sufren las carreras, el cual resulta ser uno de los problemas mas críticos por la magnitud de los impactos negativos en términos de despilfarro de recursos escasos, pérdidas de tiempo y frustraciones como perjuicios fundamentales para los estudiantes [1]. Un ejemplo más; estudiar los factores que pueden ocasionar la brecha entre las duraciones teóricas y reales de las diferentes carreras. La minería de datos permite analizar estos fenómenos aportando nuevos conocimientos, nuevas perspectivas de los mismos mediante la aplicación combinada de conceptos de campos consolidados sobre la información existente.

En cuanto a las investigaciones adelantadas sobre el tema en Colombia, se llevó a cabo en el 2005 en Bogotá el "Primer Congreso Internacional Sobre Calidad en la Educación, Repitencia, Deserción y Bajo Rendimiento Académico", donde como referencia del problema se mostraron, en términos de eficacia de la titulación los estudios llevados a cabo en varios países como Argentina, México, Chile y Costa Rica. Para Argentina se mostró que el 12% de los estudiantes que ingresan al sistema educativo se gradúa y que el mayor porcentaje de la deserción (50%), se presenta durante los tres primeros semestres académicos. Para Méjico, se señala que el 20% de los estudiantes que ingresan a un programa académico se gradúa, mientras que el 60% de los estudiantes logran completar su ciclo completo hasta la terminación de materias cinco años después del tiempo estipulado. Para Chile se estimó que el 39 % de los estudiantes que ingresan al sistema educativo se gradúa en el tiempo estimado, mientras que Costa rica exhibe un porcentaje, alto en la región del 49%.[18]

En Colombia las investigaciones sobre la deserción universitaria, por parte del gobierno no han sido muchas. Sin embargo a partir de el "Encuentro Internacional sobre Deserción en Educación Superior: Experiencias significativas", realizado por el Ministerio de Educación Nacional, el Ministerio de Educación da importancia al fenómeno de la deserción universitaria, a partir de investigaciones que permitan comprender a profundidad dicho fenómeno, a fin de diseñar políticas.

La Universidad de los Llanos realizó un estudio para doce programas académicos de pregrado. Allí las evidencias mostraron cada una de las variables incidentes en este fenómeno y se concluyó ratificando que la más alta deserción se presenta en los cinco primeros semestres del ciclo académico, donde la principal causa son

[18] Rodríguez Gama Álvaro. Compilador de las memorias del "Primer Congreso Internacional Sobre Calidad en la Educación, Repitencia, Deserción y Bajo Rendimiento Académico", Un escenario par analizar las causas y proponer soluciones. Bogotá, Septiembre 1 y 2 de 2005 [23]

motivos académicos, siendo que más afecta el bajo rendimiento debido a la repitencia de las asignaturas. En cuanto a la retención, el 56% de los estudiantes finaliza su plan de estudios. El estudio no encontró correlación entre el Puntaje de la prueba de estado y el reempeño académico de los desertores, pero sí encontró, por causas académicas, relación entre el hecho de no recibir orientación vocacional y las altas tasas de deserción inicial y temprana. En cuanto a las causas socioeconómicas, la Universidad de los Llanos halló una fuerte incidencia de los problemas económicos y de generación de ingresos en relación con la deserción, aunque en el contexto general de la institución no es la principal causa.[19]

En la costa Atlántica, la Corporación Educativa Mayor del Desarrollo Simón Bolívar de Barranquilla, en su estudio de deserción estudiantil en el programa de psicología, se concentró en la descripción de factores educativos, personales y económicos utilizando herramientas de análisis estadístico. Entre las conclusiones se vio que el factor económico es el que más incide para la recurrencia de este fenómeno, soportado por el hecho, que el 66% de los desertores están dentro de esta categoría. Sin embargo, se evidencian otros factores que influyen el proceso entre los cuales se encuentran: factores personales como enfermedad, embarazo, accidentes y matrimonio y separación; factores familiares que involucran variables de cambio de residencia, separación de sus padres y calamidades domésticas; factores académicos que incluyen eventos de pérdida de asignaturas, cambio de carrera o universidad, inseguridad en la elección de la carrera, inconformidad con la institución; factores psicológicos que involucran problemas interpersonales, emocionales, de motivación hacia el estudio y de no relación con el entorno.[20]

En Cali, La Pontificia Universidad Javeriana, los resultados del estudio de deserción para el programa de Economía[21], muestran que el apoyo familiar y el rendimiento académico previo sobre todo en matemáticas y lenguaje, son los factores que más afectan, pero que factores como sexo y número de créditos matriculados también tienen alta incidencia. El estudio califica a los tres primeros semestre como periodo crítico, en el cual la posibilidad de abandonar sus estudios es del 91% el total de deserciones voluntarias, mientras que la deserción involuntaria se ha estimado en un 95%. Las causas voluntarias que señala el estudio en mención, obedecen al proceso de adaptación al sistema académico y social de la Universidad y a las expectativas frente al programa académico que inician los estudiantes que, por ser exigente, impide cumplir con las condiciones

[19] Malagón Escobar Luz Miriam, Calderón Cañón Cesar Augusto, Soto Hernández Edwin Leonardo. Estudio de la deserción estudiantil de los programas de pregado de la Universidad de los Llanos (1998-2004), Universidad de los Llanos. Departamento de Proyección Social, Villavicencio, enero 2006 Universidad de los Llanos. Departamento de Proyección Social [24]

[20] Reyes Ruiz Lizeth. La deserción estudiantil en el programa de psicología de la Corporación Educativa Mayor Del Desarrollo Simón Bolívar, Corporación Educativa Mayor del Desarrollo Simón Bolívar. Unidad de Autoevaluación del Programa de Psicología. Barranquilla, Colombia, 2000 [25]

[21] Girón Cruz Luís Eduardo . González Gómez Daniel Enrique. Determinantes del rendimiento académico y la deserción estudiantil, en el programa de Economía de la Pontificia Universidad Javeriana de Cali. Este artículo es producto de la investigación «Determinantes del rendimiento académico y la deserción estudiantil en el programa de Economía de la Pontificia Universidad Javeriana de Cali», financiada por la Coordinación Institucional de Investigaciones, adscrita a la Vicerrectoría Académica. [26]

de permanencia. En sus conclusiones, el estudio demuestra que los principales motivos de deserción en su orden son: vocacionales (36.4%), personales (30.3%), institucionales (15.1%), económicos (9.1%) y académicos (9.1%), por lo que se identifica la deserción como un fenómeno multivariable, en el cual intervienen cuatro factores: lo individual, lo académico, lo socioeconómico y lo institucional[22]. El proceso realizado esta soportado por el uso de Herramientas estadísticas y el resultado muestra un reflejo en otra ciudad de condiciones similares en la misma universidad con sede en Bogotá.

En Bogotá, en el estudio que la Facultad de Economía de la Universidad del Rosario,[23] con herramientas de análisis estadístico, obtuvo como conclusiones que las principales causas de deserción se relacionaban con motivos académicos por el incumplimiento de los requisitos establecidos por la Universidad para mantener el cupo (mínimo promedio exigido, atraso en el plan de estudios y retraso en la culminación de materias), y no a retiros voluntarios o de naturaleza económica.

La Universidad Javeriana de Bogotá, lleva a acabo un programa de tutorías que ya cumple 10 años. Fue seleccionado dentro de la convocatoria de experiencias exitosas en la disminución de la deserción porque permite conocer las causas del abandono escolar desde la raíz misma. El tutor vela por la formación del estudiante desde las competencias ético – formativas, disciplinar y comunicativa. Otro elemento que ha incorporado es la idea de tener contacto con los orientadores profesionales de colegios y padres de familia para que desde que están en el colegio los estudiantes opten por la carrera según sus habilidades y así no se queden a mitad de camino. Este caso en concreto muestra como el problema empieza a ser enfrentado proponiendo políticas que ayuden al estudiante en su proceso universitario, disminuyendo las dificultades originadas por una deficiente formación secundaria.

En la Universidad Pedagógica Nacional se ha adelantado desde 1985 un estudio de la deserción estudiantil[24], en busca las causas reales de la deserción seleccionado como experiencia reconocida porque "arroja una concepción integral a la problemática fundamentada en estudios cualitativos y cuantitativos". Los resultados obtenidos plantean que el 50 % de estudiantes que ingresa no termina las carreras, que un 26 % de los desertores parciales – los que van y vuelven – lo hacen por causas como el embarazo y que hay dos momentos críticos en el proceso: el comienzo y el final. Las dificultades económicas son la causa más importante de la deserción en la Universidad Pedagógica y afecta a más de la tercera parte de los desertores.

---

[22] Citado por Girón y González en: Castaño *et al.* 2004, CEDE. 2005

[23] Universidad del Rosario. Facultad de Economía. Informe sobre movimiento y deserción estudiantil del pregrado de economía - primer semestre de 2001 a segundo semestre de 2004. Bogotá, 2004 [27]

[24] http://www.colombiaaprende.edu.co/html/directivos/1598/article-80793.html. Consulta por internet el 15 de agosto de 2006.[30]

Para esta universidad los motivos de la deserción antes de terminar la carrera son fáciles de detectar porque por lo general se trata de que consiguen trabajo y no se gradúan. Al comenzar la carrera, las razones pueden variar según el grupo específico, dependiendo de si los estudiantes vienen de afuera de Bogotá muchas veces no aguantan el primer semestre porque no se adaptan a la ciudad y a vivir solos. En cuanto a la gestión académica para esta universidad, como se ha detectado que el 35 % de los estudiantes mencionan causas de la deserción relacionadas con la vocacionalidad – no era lo que esperaban o no se sienten capaces – hay alternativas como tutorías, monitorías académicas, entre otras para enfrentar la deserción y disminuirla. De igual forma, Bienestar Universitario hace parte de las soluciones brindando servicio médico y psicológico gratuito, escuela maternal para estudiantes que son padres y servicio odontológico. En cuanto a la dimensión pedagógica motivan la creación de redes de apoyo mutuo a estudiantes a través de grupos focales, talleres y plenarias para promover espacios de reflexión frente a las inconformidades con la Universidad. A raíz del estudio, surgió el Centro de Acompañamiento Académico de Estudiantes (COAE) en donde se gestan proyectos para darle solución a los casos que detectan. La situación aquí planteada como reflejo de una institución pública, indica que el problema no es solamente para las universidades privadas, pero sí reflejan una condición particular relacionada con la idiosincracia de sus estudiantes y nivel socioeconómico y geográfico de los cuales proceden.

En la universidad para la que se va a desarrollar el prototipo, se realizó un estudio adelantado como tema de maestría en Dirección universitaria de la Universidad de los Andes, relacionado con deserción Universitaria[25]. Su autor no esta de acuerdo con la definición del término desertor, pues lo califica dentro de un entorno castrense y que recae sobre el estudiante toda la culpa de dicha situación. Más bien depende de elementos conocidos antes, durante y después no solo relacionados con el estudiante, sino con la universidad y su entorno. El estudio netamente estadístico, usando herramientas como SAS y Microsoft Excel, tuvo como universo una población de 800 estudiantes que abandonaron de la cohorte 1990 – 1995; una población de estudio de 170 estudiantes y fuente de información la encuesta directa. Se escogieron 49 variables relacionadas con el entorno socio-familiar, sucesos académicos, factores económicos, personales, culturales-ambientales e institucionales y cuyos resultados permitieron identificar al estudiante de ésta universidad.

La Universidad de de Los Andes, se vinculó al proceso apoyando al ministerio de Educación Nacional, realizando un estudio de deserción[26] con la información de cuatro universidades públicas, identificó las principales variables relacionadas con la deserción y se realizó una evaluación del impacto de las estrategias implementadas por las Instituciones de Educación Superior (IES). Posteriormente

[25] Álvarez Manrique, José María. Etiología de un sueño o el abandono de la universidad por parte de los estudiantes por factores no académicos. Tema de Maestría en dirección Universitaria, Universidad de los Andes. 1996.[9]

[26] Universidad de los Andes. CEDE. Deserción en las instituciones de educación superior en Colombia. [28]

se han vinculado 33 universidades caracterizadas como las que más presentan el fenómeno de deserción. Para ello desarrolló una herramienta para realizar su medición, el Spadies (Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior) con información de 790.000 estudiantes. La herramienta está soportada en otra herramienta estadística llamada STATA.

El estudio evaluó factores que predicen la deserción, tales como el académico, el socioeconómico, los institucionales (como el programa en el que está inscrito y los apoyos financieros o académicos que recibe), el género del estudiante, su edad, el número de hermanos que tiene y el puesto que ocupa en la familia. Además, establece que la deserción se debe mirar por cohorte, es decir, el comportamiento de un conjunto de estudiantes que ingresan en un semestre en particular. Como premisa considera desertor a aquel estudiante que no está matriculado durante dos o más semestres consecutivos en la misma institución. [27]

Identificó las variables asociadas al fenómeno, el cálculo del riesgo de deserción de cada estudiante para facilitar el seguimiento y evaluación de impacto de estrategias orientadas a disminuir dicho riesgo, como principales determinantes de la deserción y realizó una evaluación del impacto de las estrategias implementadas por las Instituciones de Educación Superior (IES)[28].

Las variables determinadas fueron la tasa de repitencia (materias perdidas y repetidas), desempleo de los padres, situación familiar, si estaba trabajando cuando presentó las pruebas ICFES o si sus resultados en esta prueba fueron muy bajos.

El proceso realizado se hizo en tres fases:

- El Ministerio de Educación realizó un seguimiento del fenómeno por programa y carrera.
- A nivel local se articuló la información sobre quienes tienen apoyo económico (préstamo del ICETEX o de un banco o cuántos trabajan).
- Desde las mismas instituciones se observa la evolución de los estudiantes en cada cohorte.

En los resultados obtenidos durante el período 1998–2004, se encontró que el riesgo de deserción es mayor en los primeros semestres evidenciado resultados que demuestran que el 80% de los estudiantes permanece en cada una de las cohortes al culminar el segundo semestre; en quinto semestre, ese promedio se reduce al 60%, y comienza a disminuir hasta llegar a un 44% en décimo semestre.[29]

---

[27] Tomado de: http://www.dinero.com/wf_InfoArticulo.aspx?idArt=28020 el 30 de junio de 2006.

[28] Universidad de Los Andes - CEDE. Deserción en las instituciones de educación superior en Colombia [28]

[29] MINISTERIO DE EDUCACIÓN. Acceder para quedarse: Cobertura con permanencia BOLETÍN INFORMATIVO Nº 6. Enero – marzo, Colombia, 2006 [29]

Otras de las conclusiones del estudio realizado son:
La edad promedio de ingreso a las IES, oficiales y privadas, es de 17 años. Edad que se mantiene para los diferentes programas de formación

La mayoría de los estudiantes que son admitidos en IES Oficiales provienen de familias de bajos ingresos (70%) según se muestra en la Tabla 1.

Ingreso Familiar	**Oficial (%)**	**Privada (%)**
Bajo	70	40
Medio	9	23
Alto	11	37

Tabla. 1. Estudiantes admitidos por ingreso familiar. Fuente: Muestra IES. Cálculos del CEDE.

El 52% de los estudiantes de la muestra asisten a IES Oficiales y el 48% restante a IES privadas.
El 90% de los estudiantes asiste a programas de formación Universitaria.

La mayoría de los admitidos en las IES, oficiales y privadas, son estudiantes que obtuvieron puntaje del ICFES Bajo y Medio, sin embargo, ese porcentaje es más alto entre las IES privadas. Ver tabla 2.

Nivel de Clasificación del Puntaje del ICFES	**Oficial (%)**	**Privada (%)**
Bajo	36	42
Medio	36	32
Alto	28	26

Tabla 2. Nivel de clasificación de ingreso de estudiantes con base en el puntaje Icfes obtenido. Fuente: Muestra IES. Cálculos del CEDE

La tabla 3, muestra que el 13% de los estudiantes de la muestra se beneficiaron de programas de apoyo financiero

Programa %	**Estudiantes %**
Financiero	13.2
Académico	0.6
otros	0.1
Financiero y académico	0.1
Financiero y otros	0.1
Ninguno	86.0

Tabla 3. Porcentaje de estudiantes que obtuvieron algún apoyo en la realización de sus estudios. Fuente: Muestra IES. Cálculos del CEDE

La deserción es menor entre los estudiantes que reciben programas de apoyo financiero durante un período más largo.

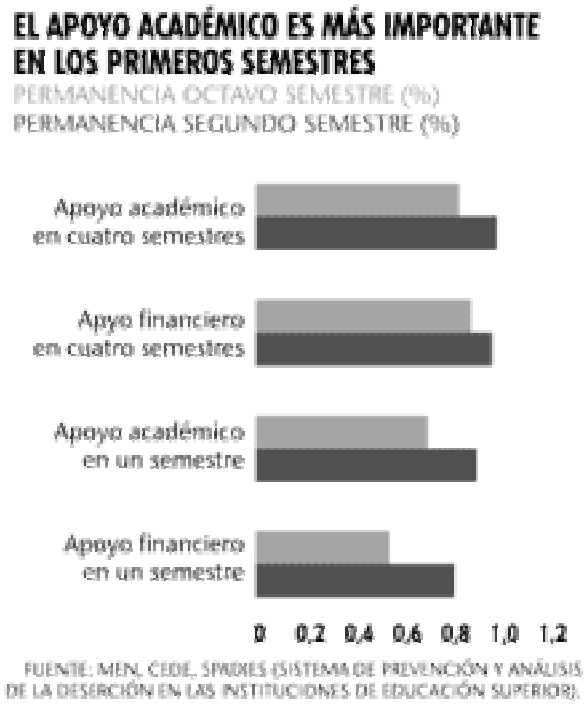

Figura 8. Efecto del apoyo académico en la retención (deserción) estudiantil universitaria.

Como estadísticas adicionales se encontró que un 20 % de los estudiantes deja las aulas por falta de dinero, un 45 % lo hace por mal rendimiento académico y un 10 % lo hace por un factor netamente vocacional. Es decir, no era lo que querían.

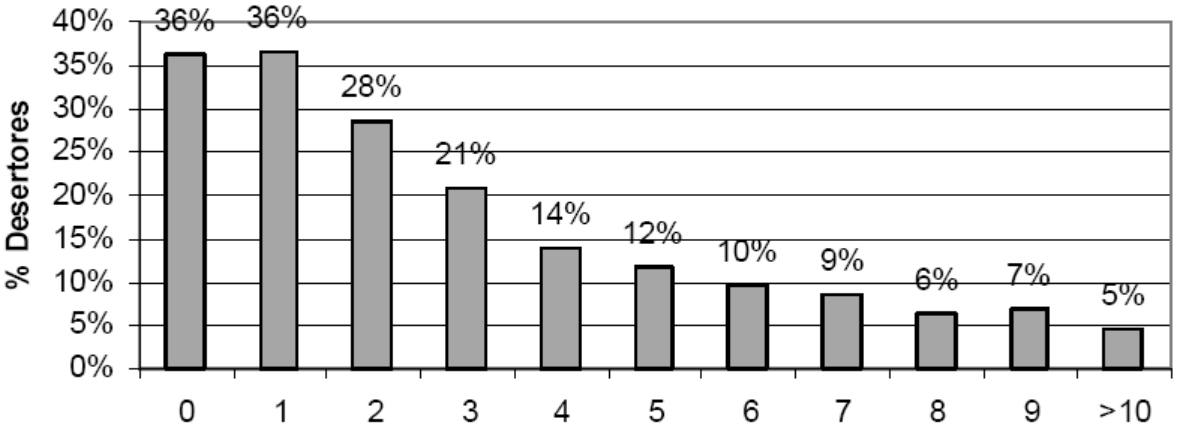

Figura 9. Fuente: Muestra IES. Cálculos del CEDE

Las condiciones determinantes para considerar a un estudiante como desertor o no (estudiante activo) se muestran en los ejemplos tabulados en la tabla 4. Aquí se indica cómo un estudiante que pudo estar estudiando por dos semestres consecutivos en los semestres 1999-1 y 1999-2 no volvió (ejemplo 1) se considera desertor, o como también se puede catalogar como desertor a un estudiante que no ha regresado en los semestres 2004-1 y 2004-2, es decir

después de un año. Las condiciones planteadas en cada uno de los ejemplos de la tabla 4, son que el estudiante ingreso en el 1999-1.

CASO	SEMESTRES DE LA MUESTRA														CLASIFICACIÓN
	1998-1	1998-2	1999-1	1999-2	2000-1	2000-2	2001-1	2001-2	2002-1	2002-2	2003-1	2003-2	2004-1	2004-2	
Ejemplo 1	x	x													Desertor
Ejemplo 2			x	x			x	x	x	x	x				Desertor
Ejemplo 3			x	x			x	x	x	x	x	x			Desertor
Ejemplo 4			x				x				x				Desertor
Ejemplo 5			x												Desertor
Ejemplo 6		x	x	x	x	x	x	x							Desertor
Ejemplo 7												x			Desertor
Ejemplo 8			x	x			x	x	x	Graduó					No desertor
Ejemplo 9	x	x					x	x	x	x	x	x	x	Graduó	No desertor
Ejemplo 10			x	x			x	x	x	x	x	x	x		Censurado

Tabla 4. Casos de muestra para clasificar o no a un estudiante como desertor. Fuente: Muestra IES. Cálculos del CEDE

El mismo estudio estableció la figura 9, en la cual se muestra una curva de supervivencia, para dos tipos de universidad: públicas y privadas. Allí se muestra que el comportamiento es muy similar. Aunque las universidades publicas tengan un menor porcentaje de deserción, en cada uno de los semestres de una misma cohorte y que la diferencia porcentual entre los semestres 3 al 7 es mayor, reduciéndose esta diferencia a partir del 8 semestre hasta ser muy similares o muy poca la diferencia a partir del 9 semestre.

.

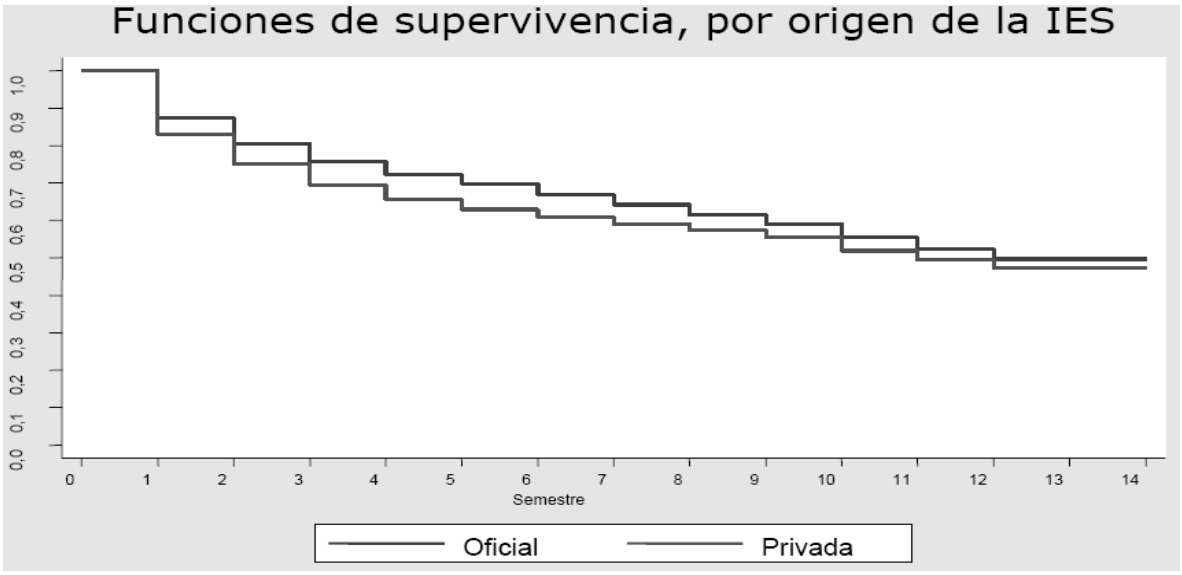

Figura 10. Curva de retención universitaria por origen de tipo de universidad: privada o pública. Fuente: Muestra IES. Cálculos del CEDE

9 ELABORACION DEL PROTOTIPO. DESARROLLO DE PROCESO.

9.1 Conocimiento de la Organización.

La Universidad en estudio, se tipifica como Universidad de la franja media, clasificación relacionada con el valor de la matricula semestral que un estudiante paga en su carrera. La población estudiantil es en su gran mayoría proveniente de los estratos socio-económicos 2 y 3 y en muy pequeño porcentaje de estrato 1 y 4 que vienen de todas las regiones del país. También se caracterizan porque no pudieron entrar a una universidad pública y que tampoco cuentan con el dinero suficiente para ingresar a otra universidad de mayor costo por semestre. Sin embargo, lo anterior no significa que el costo por semestre sea el más bajo de la franja ya que al contrario está entre las de más alto costo de dicha franja.

Actualmente, al igual que muchas otras universidades, está realizando el proceso de acreditación institucional frente al ICFES, de la mayoría de los programas de pregrado diumos y nocturnos que ofrece a la comunidad y cuenta ya con el registro calificado de todos ellos. También ofrece programas posgraduales a nivel de especialización en distintas áreas del conocimiento. La cantidad de estudiantes que actualmente están matriculados es de aproximadamente 7000, distribuidos en las facultades de Derecho, Ingenierías, Ciencias Económicas y Contables y Administración. También como ya se mencionó, esta universidad está interesada en conocer más detalladamente qué esta sucediendo con los estudiantes, pues su índice de deserción es alto en comparación con las otras universidades de la franja media. Estos datos se obtuvieron de estudios estadísticos de diversos departamentos de la universidad y de un estudio de investigación realizado por un fundador de la universidad. [30]

La población seleccionada para el prototipo dentro de la universidad fue la facultad de sistemas, con datos tomados desde el primer periodo del año 2000 hasta el primer periodo del año 2006-1, tanto estudiantes diumos como estudiantes nocturnos. El proceso llevado a cabo se detalla en las secciones del presente capìtulo.

9.2 Plan del proyecto

Para llevar el proyecto adelante, es necesario definir las etapas que lo integran. Esto es esencial para el éxito del proceso de inteligencia de negocios (BI). Considerando las principales propuestas existentes, se puede adoptar como guía para el proyecto el modelo de proceso propuesto por CRISP-DM en su versión 1.0 [8]. Éste delinea una secuencia de etapas a realizar (seis, más precisamente) y además detalla directivas paso a paso, tareas y objetivos a lograr en cada etapa del proceso. Como ventajas tiene: ser de distribución libre (permitiendo gozar de un abanico de recursos a utilizar), independiente de la aplicación y generación de producto informático acorde a las necesidades de cada organización.

[30] Álvarez Manrique, José Maria. Etiología de un sueño o el abandono de la universidad por parte de los estudiantes por factores no académicos. Tesis de maestría Universidad de los Andes. 1996 [9]

Analizando la propuesta y cotejando con el contexto del proyecto se decidió establecer una variante de la misma, complementándola con la metodología propuesta por Berry – Linoff [11] :

9.3 Establecer la población objetivo.

Definida la metodología a usar el siguiente paso fue determinar la población objetivo. En este caso se optó por el departamento de Sistemas de la facultad de Ingeniería), tomando como referencia las siguientes situaciones:

Estudiantes que ya no están en la universidad:

- A este grupo pertenecen los egresados. Muestra representativa para el estudio de retención.
- Los no egresados pero que por alguna razón ya no están en la universidad.
  - ✓ En este grupo están los que salieron por razones económicas, de trabajo, movilidad, o cambio de universidad, no continuaron los estudios, perdieron el semestre y no continuaron (retiro voluntario).
  - ✓ Los que fueron sancionados por la universidad por condiciones disciplinarias en contra del reglamento: promedio general de notas, pérdida de una asignatura por tercera vez, agresión, delincuencia. (retiro involuntario).

Los que todavía están.

- Estudiantes de los últimos dos semestres. Muestra representativa para el estudio de retención.
- Estudiantes de primer, segundo y tercer semestre que permiten establecer el proceso de deserción de los primeros semestres cuando esta tiene su mayor volumen y que permitirá entrenar al algoritmo de clasificación en cuanto a cuántos estudiantes de un semestre continuarán el siguiente semestre.

Adicionalmente los estudiantes de primer semestre al séptimo semestre, servirán de muestra representativa para la aplicación de resultados, análisis de retención (curvas de sobrevivencia,y curvas de retención) y análisis de amenazas. Para ello se eligen aquellos estudiantes que ingresaron en uno de estos periodos y se le hace seguimiento en los semestres subsiguientes para conocer su continuidad o no dentro del sistema, tomando lapsos de tiempo de 5 años (cohorte) para estudiantes diurnos y 5.5 años para noctumos, calculados a partir de 2001-I a 2005- II para estudiantes diurnos y de 2001-I a 2006-I nocturnos. El comportamiento general se hace sumando la cantidad de estudiantes que ingresan al primer semestre y de ellos cuales siguen en el segundo y así sucesivamente hasta el décimo u onceavo semestre. Los índices generales de variación relativa se calculan tomando el total de un semestre n+1 comparado con el total de un semestre n. Índices que permiten obtener la curva de retención general de la universidad en estudio.

9.4 Requerimientos de Minería de Datos

Para realizar un proyecto que involucre la Inteligencia de Negocios, con herramientas de minería de datos es necesario tener en cuenta dos consideraciones fundamentales: El (los) objetivo(s) del negocio y el (los) objetivo(s) de la minería de datos.

Un objetivo de negocio expresa los objetivos en terminología del negocio. Por ejemplo, un objetivo de negocio es “Determinar la planta docente necesaria para cubrir la cantidad de cursos (asignaturas) a ofrecer en un determinado periodo lectivo” o “Comprender porque muchos estudiantes tienen notas inferiores al promedio en determinadas materias”.

Un objetivo de Minería de Datos expresa los objetivos del proyecto en términos técnicos. Para el ejemplo anterior, un objetivo de Minería de Datos es “predecir cuales estudiantes continuaran su carrera en forma normal” o “clasificar el rendimiento de los alumnos en determinada asignatura”.

Para plantear las soluciones al problema de retención y deserción estudiantil en la universidad en estudio es necesario encontrar la razón o razones por la(s) cual(es) se están retirando los estudiantes, para fortalecer aquellas áreas que motivan a que los estudiantes se queden o hacerles la oferta correspondiente para que no se retiren.

Por lo tanto lo fundamental es formular los problemas de negocios de manera que ellos puedan ser tratados a través de minería de datos. En general un requerimiento o problema de negocio puede ser solucionado por requerimientos de Minería de Datos.

Siguiendo la metodología establecida, el primer objetivo del analista de los datos es entender, desde una perspectiva de negocios, lo que el usuario desea realmente lograr. En el presente trabajo, los requerimientos fueron impulsados fundamentalmente por el autor del proyecto trabajando en colaboración con los usuarios finales, personas de áreas de la Universidad ligadas a la toma de decisiones, tales como: Presidencia, vicepresidencia, y Bienestar Universitario, principalmente porque son las responsables del tratamiento de los asuntos referentes a los alumnos; temática que más se conoce por parte del grupo y para la cual era más fácil enfocar propuestas y sugerencias.

La elaboración de los requerimientos u objetivos fue compartida con el departamento de Sistemas de la Facultad Ingeniería y de manera indirecta (vía documentación existente) con el grupo encargado del proceso de Acreditación ante el CNA.

La aplicación de la metodología se llevo a cabo según las etapas descritas a continuación.

9.4.1 Etapa 1 - Tareas realizadas durante la Comprensión de los datos.

Definición de fuente de datos:

Durante el proyecto se trabajó con datos provenientes de las siguientes fuentes:

- Base de datos operacional de estudiantes con extracción correspondiente a los años 2001- I a 2006 -I.

- Información existente de estudiantes en Admisiones, Registro y control y en Cartera (particularmente) la información sobre préstamos existentes de los estudiantes con el ICETEX.

- Encuestas aplicadas a los estudiantes de primer semestre, encuestas aplicadas a estudiantes de 2do a 9no semestre y encuestas aplicadas a estudiantes de 10mo y 11avo semestre.

- Encuestas realizadas con motivo de las condiciones para el proceso de acreditación, para todos los estudiantes activos.

- Encuestas realizadas por la asociación de egresados, donde se encuentra la información aportada por los estudiantes egresados.

- Encuestas realizados por bienestar universitario a los estudiantes que recién ingresan a la universidad.

Importación de los datos:

Para poder trabajar con estas fuentes de datos se optó por importarlas a una base de datos en Oracle, para realizar las tareas de pre-procesamiento y análisis de los datos sin afectar el funcionamiento del resto de los sistemas.

Descripción de los datos:

Como parte de esta tarea, se realizaron las actividades referentes a la comprensión de la estructura y significado de los datos. Ante la existencia de una escasa documentación de la base de datos, se tuvo que realizar un análisis previo con el fin de poder discernir el significado de las tablas y los códigos de algunos atributos para luego poder generar una documentación que permitiera avanzar con el proyecto. Este análisis fue realizado con base en las entrevistas con el personal a cargo y de la base de datos operacional. Si bien el análisis no fue exhaustivo implicó un considerable tiempo puesto que se trata de un proceso de ingeniería inversa.

Exploración de los datos:

Esta tarea se realizó trabajando con diferentes utilidades graficas de visualización que trae incorporadas el Oracle Data Miner. El objetivo fue poder entender el contenido de los datos y realizar una valoración inicial de la calidad de los mismos.

Se realizó una revisión cuidadosa de los datos en lo referente a las condiciones que los mismos deben cumplir. En primer lugar se revisó, que existieran datos que a priori se consideraron como candidatos de participar en los diferentes procesos. Por lo tanto se realizaron tareas previas exhaustivas de análisis de los datos y la posterior generación de diferentes reportes en cuanto a la actualización necesaria e incorporación de nuevos datos durante un tiempo prudencial. En segundo lugar, se consideró que se debe tener niveles mínimos de calidad, es decir que haya la menor cantidad de inconsistencias y que no exista incoherencia en cuanto a los valores de los mismos (por ejemplo, tres valores para el campo sexo), etc.

Para disminuir el riesgo de la calidad de los datos, se realizó el análisis utilizando la documentación generada y mediante actividades exploratorias en el sitio donde estaban ubicados dichos datos. Finalmente se desarrolló el proyecto, dado que existían altas probabilidades de alcanzar los resultados previstos.

9.4.2 Etapa 2 - Preparación de los datos

En esta etapa se desarrollaron las actividades de construcción del dataset (datos de entrada para los distintos algoritmos). Las tareas de preparación de datos se realizan generalmente varias veces durante el proceso de Minería de Datos, puesto que a medida que se descubre información suele ser necesario cambiar o agregar nuevos datos para considerar en el análisis. Esta etapa incluye selección de tablas, registros y atributos, transformación y limpieza de datos. La forma de implementar cada tarea depende de las funcionalidades que brinda la herramienta de Minería de Datos. En esta etapa se realizaron las siguientes tareas:

- Cargar datos desde el período académico 2001-1, para las variables mostradas en la tabla 5, complementándola con las fuentes de datos ya citadas. Este proceso se ejecuto varias veces, debido a la necesidad de sacar a las variables que no aportan mucho al modelo, pero que influyen debido a la cantidad de datos con los que cuentan.

- Ejecutar el algoritmo de puntuación de atributos, para establecer cuáles son las variables que más peso tienen dentro del grupo de datos (Dataset) y verificar sí ésta es relevante o no para los distintos modelos.

- Utilización del algoritmo de árboles de decisión para la extracción de la reglas tanto de deserción como de retención según se ve en la figura 1. y la validación correspondiente por los usuarios.

- Utilización de los algoritmos para el agrupamiento de los datos (clustering), tomando como referencia que en gran parte debe ser utilizando variables categóricas, para evitar gran diversidad de los datos. Igualmente se debe elegir el valor de K para el algoritmo de K-means.

- Analizar y reacondicionar las reglas del árbol de decisión de acuerdo a los resultados.

Variables para la generación del cluster de datos. Construcción del dataset:

Las variables que se muestran a continuación son producto de la integración de diferentes estudios como el relacionado en [12] [13] y resultado de las encuestas que se están aplicando en diferentes momentos de la vida del estudiante, tales como al momento de la inducción en el primer semestre; conocimiento de la percepción del estudiante de su ambiente y condición a partir del segundo semestre hasta el 9 y finalmente la encuesta de egresados aplicada a los estudiantes que terminan su ultimo semestre académico en la universidad. Estas encuestas estructuradas están orientadas a determinar el valor de las variables económicas, sociales, personales, culturales, académicas del bachillerato, académicas de la universidad, que permiten posteriormente conocer más detalladamente la situación de cada estudiante en particular y de todo un grupo de estudiantes (por semestres, carreras, facultades y toda la universidad) en general. Las variables propuestas inicialmente se muestran en la tabla 5 y basadas en el estudio de incidencia que se quiere atacar, para conocer el problema desde diferentes perspectivas.

La funcionalidad esperada en la elección de estas variables es que puedan ser aplicadas a todos los estudiantes independientemente de sí son estudiantes que terminaron o son estudiantes retirados. Durante el proceso, dependiendo de la calidad de los datos, es decir si tienen en un alto porcentaje información, se van seleccionando las variables que quedan o variables que dada su pertinencia y según el algoritmo de puntuación de atributos, no quedan en los modelos a realizar.

Los datos nulos existentes en algunas de las variables, se tratan en dónde es posible de la siguiente forma: para variables categóricas se usa la Moda y para variables numéricas se remplaza por el promedio. Se categorizarón la mayoría de las variables debido a la necesidad de disminuir la cantidad de grupos que pueden existir, en caso de variables con datos con gran espectro.

Para cada requerimiento es necesario construir el dataset con el contenido de los datos a analizar. El ODM facilita mucho la realización de esta tarea, ya que para cada algoritmo o prueba de un algoritmo se pueden elegir las variables que formaran parte del grupo (dataset) deseado.

Las siguientes son las variables iniciales propuestas para el dataset.

Variable	Tipo de variable
Situación personal del estudiante	
IDESTUDIANTE	
SEXO	Categórica
LUGARNACIMIENTO	
FECHANACIMIENTO	
EDAD	Transformada
DIRECCIONRESIDENCIA	Transformada
TELEFONORESIDENCIA	
ESTADOCIVIL	Categórica
NROHIJOS	
LUGAROCUPAHERMANOS	Categórica
ACTUALMENTEVIVECON	
PERTENECEAGRUPOS	Categórica
ENCONSULTASICOLOGICA	Categórica
MOTIVOCONSULTASICOLOGICA	Categórica
ENCOFLICTOSPERSONALESACUDEA	Categórica
RELACIONCONFAMILIA	Categórica
RELACIONCONAMIGOS	Categórica
RELACIONCONPAIS	Categórica
COSTEOESTUDIOS	Categórica
DEPENDENCIAECONOMICA	Categórica
PROYECTOSAUTOGESTION	Categórica
TIENECOMPUTADOR	Categórica
FORMAACCESOINTERNET	Categórica
TIENEPAPA	Categórica
TIENEMAMA	Categórica
OTROSESTUDIOS	Categórica
NROPERSONASACARGO	Categórica
MADREPADRESOLTERO	Categórica
QUIENCUIDAHIJOS	Categórica
BARRIO	Transformada
Situación laboral del estudiante	
INGRESOSMENSUALES	Transformada
TRABAJAACTUALMENTE	Categórica
HORARIOTRABAJO	Categórica
Aficiones del estudiante	
OIRMUSICA	Categórica
ACTIVIDADARTISTICA	Categórica
CINE	Categórica
TV	Categórica
BAILAR	Categórica
DEPORTE	Categórica
DORMIR	Categórica
CAMINATAS	Categórica

Variable	Tipo de variable
AMIGOSFAMILIA	Categórica
OTROS	Categórica
LEER	Categórica
Situación familiar	
DEPTORESIDEFAMILIA	Categórica
PROCEDENCIAIINGRESOFAMILIAR	Categórica
TIPOVIVIENDA	Categórica
TENENCIAVIVIENDA	Categórica
NROPERSONASCOMPARTENVIVIENDA	Transformada
EDADPAPA	Transformada
OCUPACIONPAPA	Categórica
ESCOLARIDADPAPA	Categórica
EDADMAMA	Transformada
OCUPACIONMAMA	Categórica
ESCOLARIDADMAMA	Categórica
PARENTESCO3	Categórica
EDADRELACIONADO3	Transformada
OCUPACIONRELACIONADO3	Categórica
ESCOLARIDADRELACIONADO3	Categórica
PARENTESCORELACIONADO4	Categórica
EDADRELACIONADO4	Transformada
OCUPACIONRELACIONADO4	Categórica
ESCOLARIDADRELACIONADO4	Categórica
PARENTESCO5	Categórica
EDADRELACIONADO5	Transformada
OCUPACIONRELACIONADO5	Categórica
ESCOLARIDADRELACIONADO5	Categórica
ESTRATOECONOMICO	Categórica
Estudios secundarios	
TIPOCOLEGIOTERMINO	Categórica
TITULOBACHILLER	Categórica
VALORULTIMAPENSIONCOLEGIO	Transformada
EXAMENICFES	Categórica
PUNTAJEICFES	Transformada
COLEGIOCONINTERNET	Categórica
Estudios universitarios	
JORNADA	Categórica
PROGRAMA	Categórica
SEMESTRECURSA	Categórica
TIEMPOINGRESOUNIVERSIDAD	Categórica
ACCEDEALAUNIVERSIDADPOR	Categórica
INFLUENCIAESCOGERCARRERA	Categórica
CONOCEPROGRAMAACADEMICO	Categórica
COMOCONOCIOPROGRAMA	Categórica

Variable	Tipo de variable
RELACIONCONUNIVERSIDAD	Categórica
NUMEROASIGNATURASVISTAS	Transformada
NUMEROASIGNATURASPERDIDAS	Transformada
ASIGNATURASVISTASMASDEUNAVEZ	Transformada
PROMEDIO	Transformada
ULTIMOPERIODOLECTIVOCURSADO	Categórica
PERIODOLECTIVO	Categórica
ULTIMONIVEL	Categórica
ESTADO	Categórica

Tabla 5. Variables propuestas inicialmente para poblar el grupo de datos del modelo.

En el momento del desarrollo de este trabajo existen en el SPADIES las siguientes variables, las cuales permiten confrontar con los mismos datos los resultados y por lo tanto tener un punto de validación tanto de uno como del otro modelo. Finalmente se dejo un data set con la mayoría de estas variables más otras que son importantes para el seguimiento particular del estudiante, el cual se muestra en la tabla 6.

Variable
IDESTUDIANTE
NOMBRES
SEXO
FECHANACIMIENTO
POSICIONHERMANOS
VIVIENDAPROPIA
TRABAJABA
NIVELEDUCATIVOMADRE
INGRESOSFAMILIARES
EDADENICFES
NROHERMANOS
PUNTAJEICFES
PERIODOINGRESO
PROGRAMA
MATERIASTOMADAS
MATERIASAPROBADAS
NROAPOYOSICETEX
NROAPOYOSFINANCIEROS
NROOTROSAPOYOS
PERIODOGRADO
PERIODORETIROFORZOSO
NROSEMESTRESCURSADOS
TASAREPITENCIA
ESTADOESTUDIANTE

Variable Finales
IDESTUDIANTE
NOMBRES
SEXO
FECHANACIMIENTO
POSICIONHERMANOS
VIVIENDAPROPIA
TRABAJABA
NIVELEDUCATIVOMADRE
INGRESOSFAMILIARES
EDADENICFES
NROHERMANOS
PUNTAJEICFES
PERIODOINGRESO
PROGRAMA
MATERIASTOMADAS
MATERIASAPROBADAS
NROAPOYOSICETEX
NROAPOYOSFINANCIEROS
NROOTROSAPOYOS
PERIODOGRADO
PERIODORETIROFORZOSO
NROSEMESTRESCURSADOS
TASAREPITENCIA
ESTADOESTUDIANTE

Tabla 6. Variables de SPADIES y variables finales del aplicativo.

Limpieza del dataset:
Esta tarea consiste en realizar el tratamiento de datos faltantes (missing) y con ruidos, identificar o remover valores excepcionales (outliers) y resolver inconsistencias existentes. Los scripts que se ejecutan para realizar esta tarea se muestran en el anexo A.

Transformación y creación de los datos:
Esta tarea incluye operaciones de construcción, como la producción de atributos derivados, ingresar nuevos registros, agregación y generalización de los datos, discretización y transformación de valores de atributos existentes. Aquí fue necesario categorizar muchas de las variables numéricas, debido a que forman un gran espectro sobre el cual se mueven los datos, impidiendo obtener conjuntos de datos que proporcionan resultados satisfactorios.

9.4.3 Etapa 3 - Minería de Datos de los Datos e interpretación de resultados.

En esta etapa se seleccionan y se aplican diferentes técnicas y algoritmos de modelado calibrando sus parámetros para obtener resultados óptimos.

Generalmente, hay varias técnicas para el mismo problema. Algunas técnicas tienen requerimientos específicos sobre la estructura de los datos por lo que puede ser necesario retomar a la fase de preparación de los datos.

Resultados con los Algoritmos a utilizar.
Los algoritmos a utilizar son los disponibles en la herramienta ODM de Oracle 10G R2 .
Total de la muestra 2631 casos.
Algoritmo de conglomerados (clustering).
Se ejecutó con las siguientes las variables que tienen más completitud de los datos:
IDESTUDIANTE, SEXO, PUNTAJEICFES, PERIODOINGRESO,
MATERIASTOMADAS, MATERIASAPROBADAS, NROAPOYOSICETEX
NROSEMESTRESCURSADOS, TASAREPITENCIA
ESTADOESTUDIANTE, ULTIMOPERIODOCURSADO
ESTRATO, ESTADOCIVIL, PROMEDIO
MOTIVORETIRO

Resultados:

													%Part	%Acu
1	2631													
	2	1003											38,1	38,1
	3	1628												
		4	861										32,7	70,8
		5	767											
			6	202									7,7	78,5
			7	565										
				8	143								5,4	84,0
				9	422									
					10	62							2,4	86,3
					11	360								
						12	80						3,0	89,4
						13	280							
							14	48					1,8	91,2
							15	232						
								16	46				1,7	92,9
								17	186					
										18	93		3,5	96,5
										19	93		3,5	100,0

Tabla 7, Conglomerados para la totalidad de registros en la tabla de trabajo (Resumen)

De la tabla 7 se puede deducir que los dos primeros cluster hojas (2 y 4) representan aproximadamente el 70% (las 2/3 partes) de toda la población, siendo el 30 % (1/3) parte restante representado por 8 nodos hojas (6, 8, 10,12, 14, 16, 18 y 19).

Para determinar las características de los conglomerados más significativos (2, 4) y los menos significativos (18, 19) se detallan a continuación en las tablas 8, 9, 10 y 11 respectivamente.

"Attribute"	"Centroid Value"
"ESTADOCIVIL"	"soltero"
"ESTADOESTUDIANTE"	"activo"
"ESTRATO"	"3"
"MATERIASAPROBADAS"	"33.406651778630824"
"MATERIASTOMADAS"	"37.58424725822529"
"MOTIVORETIRO"	"1"
"NROAPOYOSICETEX"	"2.564338235294134"
"NROSEMESTRESCURSADOS"	"6.1286141575274105"
"PERIODOINGRESO"	"2005-2"
"PROMEDIO"	"3.3747043473466007"
"SEXO"	"M "
"TASAREPITENCIA"	"0.148474576271187"
"ULTIMOPERIODOCURSADO"	"2006-1"

Tabla 8. Detalle del conglomerado 2, resultante de aplicar el algoritmo de Conglomerados.

La regla resultante es:
IF
ESTADOCIVIL in (soltero) and ESTADOESTUDIANTE in (activo) and ESTRATO in (2.0,3.0) and MATERIASAPROBADAS <= 63.0 and MATERIASAPROBADAS >= 0.0 and MATERIASTOMADAS <= 70.0 and MATERIASTOMADAS >= 1.0 and MOTIVORETIRO in (1.0) and NROAPOYOSICETEX <= 2.8 and NROAPOYOSICETEX >= 2.2 and NROSEMESTRESCURSADOS <= 11.0 and NROSEMESTRESCURSADOS >= 1.0 and PERIODOINGRESO in (2000-1, 2001-1, 2001-2, 2002-1, 2002-2, 2003-1, 2003-2, 2004-1, , , ,) and PROMEDIO <= 4.1 and PROMEDIO >= 2.46 and SEXO in (M) and TASAREPITENCIA <= 0.4 and TASAREPITENCIA >= 0.0 and ULTIMOPERIODOCURSADO in (2006-1)
THEN
Cluster equal 2
Confidence (%)=82.3529411764706
Support =826

Este conglomerado agrupa fundamentalmente a los estudiantes activos, hombres, solteros, de estratos 2 y 3, que tienen promedio académico histórico entre 2.46 y 4.1, han recibido entre 2.2 y 2.8 créditos del ICETEX y la tasa de repitencia esta entre 0 y 0.4. Las variables número de materias tomadas, número de materias aprobadas, período de ingreso, último período cursado, número de semestres cursados no aportan al modelo ya que los valores aquí definidos comprenden el rango de valores sobre los cuales se mueve cada una de esas variables; la variable motivo de retiro ocasiona problemas de ruido en el resultado, ya que un estudiante activo no puede estar retirado con código 1. El número de casos es de 1003 casos.

El grado de precisión es del 82%, probabilidad que indica que un estudiante con las anteriores características esté en este grupo. Este conglomerado permite identificar el tipo de estudiante que probablemente se quedará a terminar su carrera. En este caso se debe profundizar en el conocimiento y caracterización, agregando nuevas variables que permitan entender el por qué el estudiante sigue en la universidad. Estas variables deben estar relacionadas con sus características personales, del entorno familiar y social, como también académicas del bachillerato, de las cuales para el estado actual del proyecto aún no se tienen completamente digitadas.

Para el cluster 4.

"Attribute"	"Centroid Value"
"ESTADOCIVIL"	"soltero"
"ESTADOESTUDIANTE"	"inactivos"
"ESTRATO"	"3"
"MATERIASAPROBADAS"	"9.223024103332326"
"MATERIASTOMADAS"	"12.974312563162723"
"MOTIVORETIRO"	"1"
"NROAPOYOSICETEX"	"2.526567944250886"
"NROSEMESTRESCURSADOS"	"2.28434715954513"
"PERIODOINGRESO"	"2001-2"
"PROMEDIO"	"2.820599730212533"
"SEXO"	"M "
"TASAREPITENCIA"	"0.417162630422054"
"ULTIMOPERIODOCURSADO"	"2004-2"

Tabla 9. Detalle del conglomerado 2, resultante de aplicar el algoritmo de Conglomerados.

La regla resultante para este cluster es:
IF
ESTADOCIVIL in (soltero) and ESTADOESTUDIANTE in (inactivos) and ESTRATO in (3.0) and MATERIASAPROBADAS <= 25.200000000000003 and MATERIASAPROBADAS >= 0.0 and MATERIASTOMADAS <= 28.6 and MATERIASTOMADAS >= 1.0 and MOTIVORETIRO in (1.0) and NROAPOYOSICETEX <= 2.8 and NROAPOYOSICETEX >= 2.2 and NROSEMESTRESCURSADOS <= 4.0 and NROSEMESTRESCURSADOS >= 1.0 and PERIODOINGRESO in (2000-1, 2001-1, 2001-2, 2002-1, 2002-2, 2003-1, 2003-2, 2004-1, 2005-1) and PROMEDIO <= 3.69 and PROMEDIO >= 2.0500000000000003 and SEXO in (M) and TASAREPITENCIA <= 1.0 and TASAREPITENCIA >= 0.0 and ULTIMOPERIODOCURSADO in (2000-1, 2001-1, 2001-2, 2002-1, 2002-2, 2003-1, 2003-2, 2004-1, 2004-2, 2005-1, 2006-1)

THEN
Cluster equal 4

Confidence (%)=80.1393728222996

Support =690

Este conglomerado está constituido por estudiantes hombres, solteros, Inactivos, de estrato 3, que cursaron máximo 4 semestres, vieron máximo 25 asignaturas, con promedio académico entre 2.0 y 3.69, el motivo de retiro es el 1, (problemas económicos), el número de créditos con el Icetex está entre 2.2 y 2.8. Las variables número periodo de ingreso, último período cursado, y tasa de repitencia no aportan al modelo ya que los valores aquí definidos comprenden el rango de valores sobre los cuales se mueve cada una de esas variables. El número de casos en este conglomerado es de 861. El grado de confiabilidad de este grupo es del 80%.

La característica de este conglomerado de ser de estudiantes inactivos y que la razón por la cual se van es por motivos económicos, determina para la universidad la aplicación de políticas orientadas a dar mayores facilidades de pago, otorgar becas o cualquiera otra orientada a que el estudiante pueda permanecer en la universidad, disminuyendo de esta forma la cantidad de estudiantes que se van por esta razón. Sin embargo el hecho que el conglomerado esté formado por un grupo de 861 estudiantes, también condiciona a una mayor profundización en el conocimiento de los mismos, agregando variables que permitan una mayor diferenciación (como las citadas en el conglomerado anterior, variables socioeconómicas y personales del estudiante).

Para el cluster 18:

"Attribute"	"Centroid Value"
"ESTADOCIVIL"	"soltero"
"ESTADOESTUDIANTE"	"inactivos"
"ESTRATO"	"3"
"MATERIASAPROBADAS"	"4.65591397849462"
"MATERIASTOMADAS"	"8.827956989247314"
"MOTIVORETIRO"	"1"
"NROAPOYOSICETEX"	"2.531249999999992"
"NROSEMESTRESCURSADOS"	"1.795698924731182"
"PERIODOINGRESO"	"2000-2"
"PROMEDIO"	"3.093437967115103"
"SEXO"	"M "
"TASAREPITENCIA"	"0.539139784946236"
"ULTIMOPERIODOCURSADO"	"2000-2"

Tabla 10. Detalle del conglomerado 18, resultante de aplicar el algoritmo de Conglomerados.

La regla correspondiente es:
IF
ESTADOCIVIL in (soltero) and ESTADOESTUDIANTE in (inactivos) and ESTRATO in (3.0) and MATERIASAPROBADAS <= 12.600000000000001 and

MATERIASAPROBADAS >= 0.0 and MATERIASTOMADAS <= 14.8 and MATERIASTOMADAS >= 1.0 and MOTIVORETIRO in (1.0,4.0) and NROAPOYOSICETEX <= 2.8 and NROAPOYOSICETEX >= 2.2 and NROSEMESTRESCURSADOS <= 4.0 and NROSEMESTRESCURSADOS >= 1.0 and PERIODOINGRESO in (2000-2) and PROMEDIO <= 3.28 and PROMEDIO >= 2.87 and SEXO in (M) and TASAREPITENCIA <= 1.0 and TASAREPITENCIA >= 0.2 and ULTIMOPERIODOCURSADO in (2000-2, 2001-1)

THEN
Cluster equal 18

Confidence (%)=83.870967741 9355
Support =78

A medida que se avanza más en la cantidad de hojas del árbol se hacen mucho mas particularizadas las condiciones por las cuales se está en este conglomerado. Como resultado aquí en este conglomerado se encuentra a los estudiantes que ingresaron en el 2000-2, solteros, Inactivos, de estrato 3, que cursaron máximo 4 semestres, vieron máximo 14 asignaturas, con promedio académico entre 2.87 y 3.28, el motivo de retiro el 1 (problemas económicos) o el 4 (problemas académicos), el número de créditos con el Icetex está entre 2.2 y 2.8 y tasa de repitencia entre 0.2 y 1.0 y con último período cursado en el 2000-2 o 2001-1. El número de casos en este conglomerado es de 93. El grado de confiabilidad de este grupo es del 83%.

Este conglomerado identifica particularmente a un grupo de estudiantes que se retiraron entre el 2000-2 y 2001-1. Al contrario de los conglomerados analizados anteriormente, este conglomerado presenta una cantidad considerable de estudiantes con similares características (es decir muy homogéneas) por lo cual es necesario remitirse a ese lapso de tiempo en particular y confrontar otro tipo de variables relacionadas con las políticas aplicadas por la universidad, como por ejemplo si hubo incremento fuera de lo normal de el costo de la matrícula, las otras universidades no incrementaron el valor de la matrícula en igual porcentaje, la aplicación del reglamento académico fue particularmente fuerte, hubo cambio de pensum o condiciones de desmejoramiento en cuanto a calidad por parte de la universidad que fueron percibidos por los estudiantes, etc. Encontrar las variables o razones por las cuales este evento se llevo a cabo, permite evitar que la situación se repita o planear de otra manera la aplicación de las normas si es que es necesario su aplicación.

Para el cluster 19:

"Attribute"	"Centroid Value"
"ESTADOCIVIL"	"soltero"
"ESTADOESTUDIANTE"	"inactivos"
"ESTRATO"	"3"
"MATERIASAPROBADAS"	"13.319822235843962"
"MATERIASTOMADAS"	"17.329842200809928"

"MOTIVORETIRO"	"4"
"NROAPOYOSICETEX"	"2.503360215053758"
"NROSEMESTRESCURSADOS"	"2.72035393736274"
"PERIODOINGRESO"	"2003-2"
"PROMEDIO"	"2.845816577462753"
"SEXO"	"M "
"TASAREPITENCIA"	"0.35920732249178"
"ULTIMOPERIODOCURSADO"	"2005-1"

Tabla 11. Detalle del conglomerado 19, resultante de aplicar el algoritmo de Conglomerados.

La regla resultante es:
IF
ESTADOCIVIL in (soltero) and ESTADOESTUDIANTE in (inactivos) and ESTRATO in (3.0) and MATERIASAPROBADAS <= 31.5 and MATERIASAPROBADAS >= 0.0 and MATERIASTOMADAS <= 35.5 and MATERIASTOMADAS >= 1.0 and MOTIVORETIRO in (4.0) and NROAPOYOSICETEX <= 2.8 and NROAPOYOSICETEX >= 2.2 and NROSEMESTRESCURSADOS <= 6.0 and NROSEMESTRESCURSADOS >= 1.0 and PERIODOINGRESO in (2004-2, 2005-1) PERIODOINGRESO in (2003-1, 2003-2, 2004-1, ,) and PROMEDIO <= 4.1 and PROMEDIO >= 2.0500000000000003 and SEXO in (M) and TASAREPITENCIA <= 1.0 and TASAREPITENCIA >= 0.0 and ULTIMOPERIODOCURSADO in (2003-1, 2003-2, 2004-1, 2005-1)

THEN
Cluster equal 19
Confidence (%)=77.41935483870967
Support =72

Para este conglomerado igual que para el conglomerado 18, la particularidad está en que son estudiantes que ingresaron después del 2003-1, tienen bajo promedio académico, cursaron entre dos y tres semestres y como causa de deserción esta la 4. Adicionalmente perdieron 1 de cada tres materias tomadas. La confiabilidad de este conglomerado es del 77% y el número de casos que lo soportan es de 93.

El análisis es muy similar al aplicado al conglomerado 18, ya que son estos los dos últimos conglomerados hallados por el algoritmo, sólo que la diferencia es que está constituido por estudiantes que se retiraron por causas académicas.

Aplicación del algoritmo de clasificación por árbol de decisión.
Los resultados de este algoritmo son:
Parámetros para la ejecución:
Número de casos 2631.
Random number: 12345.
Samplig type: stratified.
Equal distribution: no
Accuracy goal: maximum average
Homogeneity metric: Gini

Maximun depth: 7
Minimum records in a node: 10.
Minimum records for a split: 20.
Number of lift quantiles: 10
ROC result.

El resultado se ve en la tabla 12 mostrada a continuación.

Arbol						Soporte	Predecido	Confiabilidad	Nro casos
1	"MATERIASTOMADAS <= 23.5"					"0.49967927"	"inactivos"	"0.80359435"	"779.0"
	10	"MATERIASTOMADAS <=5.5"				"0.1359846"	"inactivos"	"1.0"	"212.0"
	2	"MATERIASTOMADAS > 5.5"				"0.36369467"	"inactivos"	"0.73015875"	"567.0"
		3	"TASAREPITENCIA <= 0.185"			"0.11289288"	"inactivos"	"0.54545456"	"176.0"
			11	"PROMEDIO <=4.06"		"0.0853111"	"activo"	"0.5714286"	"133.0"
			12	"PROMEDIO > 4.06"		"0.011545863"	"inactivos"	"0.7777778"	"18.0"
		4	"TASAREPITENCIA > 0.185"			"0.2508018"	"inactivos"	"0.81329924"	"391.0"
			5	"MATERIASTOMADAS <=10.5"		"0.12315587"	"inactivos"	"0.8958333"	"192.0"
				13	"MATERIASTOMADAS <=7.0"	"0.033354715"	"inactivos"	"0.63461536"	"52.0"
				14	"MATERIASTOMADAS >7.0"	"0.089801155"	"inactivos"	"0.99285716"	"140.0"
			15	"MATERIASTOMADAS >10.5"		"0.12764592"	"inactivos"	"0.7336683"	"199.0"
6	Materias tomadas > 23,5					"0.38935214"	"activo"	"0.738056"	"607.0"
	7	"NROSEMESTRESCURSADOS <= 10.0"				"0.30917254"	"activo"	"0.6701245"	"482.0"
		16	"NROSEMESTRESCURSADOS <=1.5"			"0.017960232"	"inactivos"	"0.71428573"	"28.0"
		8	"NROSEMESTRESCURSADOS >1.5"			"0.29121232"	"activo"	"0.6938326"	"454.0"
			17	"PROMEDIO <= 2.83"		"0.018601667"	"activo"	"0.55172414"	"29.0"
			9	"PROMEDIO > 2.83"		"0.2366902"	"activo"	"0.8102981"	"369.0"
				18	"NROSEMESTRESCURSADOS <=8.5"	"0.19499679"	"activo"	"0.7730263"	"304.0"

Tabla 12. Resultados del algoritmo de clasificación por árbol de decisión

Este algoritmo permite establecer la probabilidad (confiabilidad) con la que un estudiante “inactivo” o “activo” sea clasificado como tal dependiendo de condiciones representadas por las variables dadas al modelo. Así por ejemplo un estudiante es clasificado “inactivo” por el nodo 13 con una probabilidad de 63.46% si el número de asignaturas vistas (tomadas) es menor o igual a 7 y tiene tasa de repitencia del 18.5% o, que sea clasificado como activo con una probabilidad del 81%, sí tiene promedio académico mayor a 2.83, el número de semestres cursados es mayor a 1 y ha tomado mas de 23 asignaturas. (Nodo 9).

El algoritmo se puede complementar, para determinar por ejemplo cuántos semestres cursará un estudiante nuevo, cuando se incluyan las variables socioeconómicas aplicando el modelo ya encontrado como resultado de la aplicación sobre un conjunto de variables ya establecidas.

Métricas para el resultado anterior:

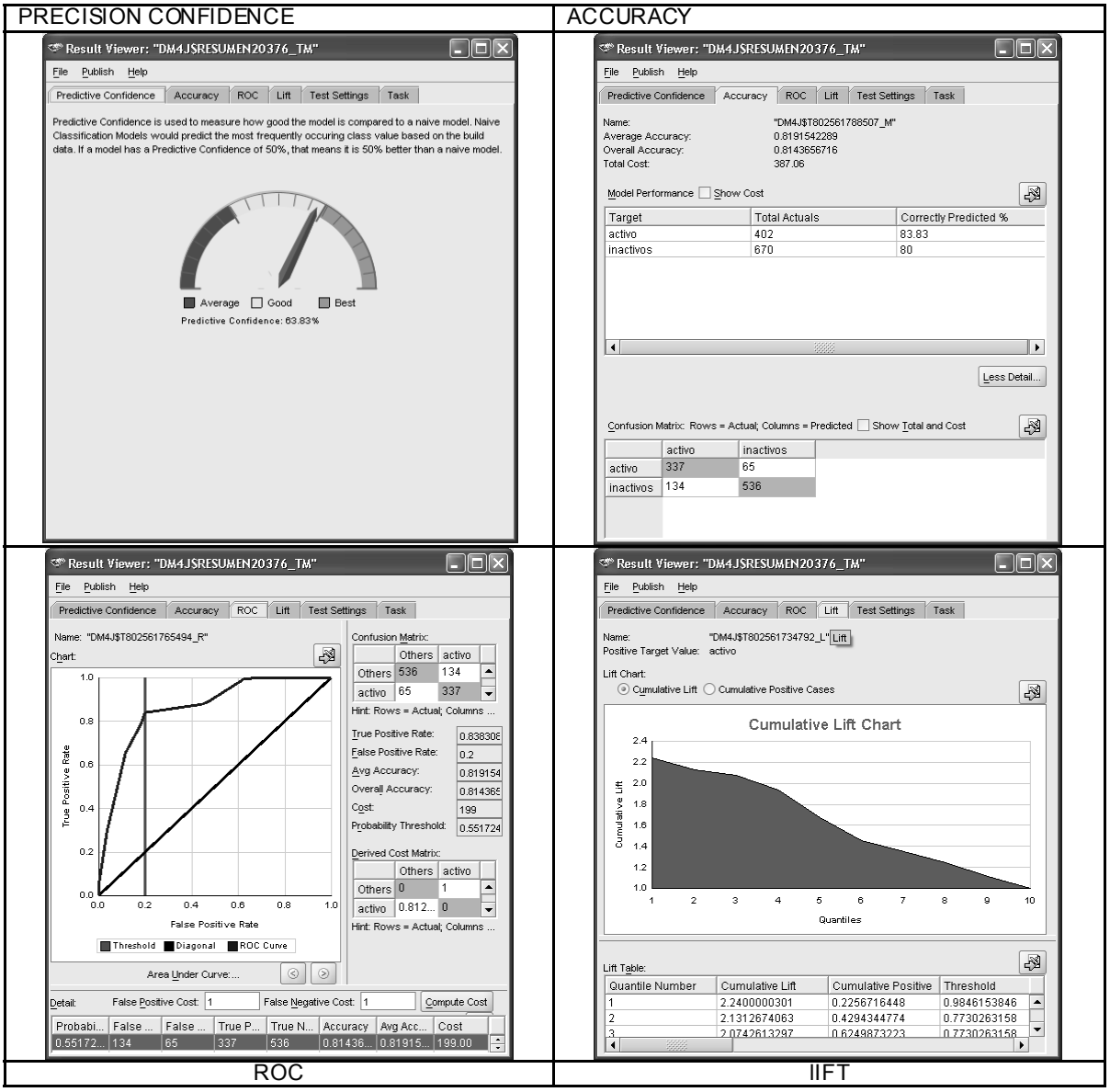

Figura 11. Métricas resultantes de la aplicación del árbol de decisión.

De las métricas anteriores se destacan: La precisión de la confiabilidad. Dada por la gráfica de la esquina superior izquierda que muestra que tan preciso es el modelo obtenido comparado con el modelo de Naive bayes. En este caso es bueno con un valor de 63.9%. La gráfica que muestra la matriz de confusión, esquina superior derecha indica que hay 134 casos que se catalogaron como activos siendo inactivos (falsos negativos) y 65 clasificados como inactivos siendo activos (falsos positivos), con los resultados mostrados en esta matriz de confusión.

La gráfica LIft (acumulado de casos positivos), mostrada en la esquina inferior derecha, muestra que para un valor positivo designado como objetivo (para el caso Estadoestudiante 'activo' o 'inactivo') los casos de prueba fueron clasificados como tales y mostrados como quantiles. La probabilidad de borde, para un quantile, es la mínima probabilidad para un objetivo positivo de ser incluido en este o cualquier quantile anterior. Lift es una medida de cuán rápido el modelo encontrara los valores objetivo positivos.

Métrica del análisis ROC (Reciever Operating Characteristics). En esta métrica el eje horizontal mide el promedio de falsos positivos como un porcentaje. El eje vertical muestra la tasa de 'verdaderos positivos'. La esquina superior izquierda es la ubicación óptima en la curva, indicando una alta tasa de verdaderos positivos vs. Falsos positivos. Esta gráfica también muestra el valor frontera del modelo el cual indica el valor aceptable entre verdaderos positivos y el valor de alarma (falsos positivos). Esta métrica es una forma de experimentar con el análisis ¿Qué pasa si el valor frontera cambia?, ¿de qué forma se afectará el modelo?

Clasificación Naive Bayes.
Parámetros.
Samplig type: startified.
Random seed: 12345.
Discretize: Quantile binning,
Categorical strategy: top N. binning.
Maximun Average Accuracy.
Los resultados se muestran en la tabla 13 a continuación.

"Attribute Name"	"Value"	"Probability"
"ESTADOCIVIL"	"soltero"	"0.92953020134228183322031143835382355990 27"
"SEXO"	"M"	"0.72483221476510027101342723968722819697 34"
"MOTIVORETIRO"	"1"	"0.68058455114822519180135981228720705097 51"
"NROAPOYOSICETEX"	"(2;4]"	"0.60869565217391308991210224072712166197 16"
"ESTRATO"	"3"	"0.52542372881355890415751639872046367675 13"
"TASAREPITENCIA"	"(;38;1]"	"0.50318471337579595857878010126656443563 9"
"MATERIASTOMADAS"	"(10;30]"	"0.47643312101910851765973101958302501057 2"
"MATERIASAPROBADAS"	"[1;401298-5;6]"	"0.45454545454545416793690192942694843529 67"
"NROSEMESTRESCURSADOS"	"(2;5]"	"0.45095541401273884475315108342083970142 91"
"NROSEMESTRESCURSADOS"	"[1;2]"	"0.41273885350318473302619368000448332952"
"MATERIASAPROBADAS"	"(6;26]"	"0.40000000000000002607341088470720527837"
"MATERIASTOMADAS"	"[1;10]"	"0.38598726114649654113753579991809334771"
"ESTRATO"	"2"	"0.36440677966101670739066880693886975425 38"
"NROAPOYOSICETEX"	"[1;2]"	"0.34782608695652165654088920606458942757 16"
"TASAREPITENCIA"	"(;08;;38]"	"0.30191082802547747950358250766950880441 64"
"PROMEDIO"	"[1	40129846432481700000000000000000E-45;2;35]"
"SEXO"	"F "	"0.27516778523489914645739772472982709145 86"
"MOTIVORETIRO"	"4"	"0.26617954070981194839617384658999400241 36"
"PROMEDIO"	"(2;35;2;88]"	"0.20819112627986328494367071918985324686 92"
"PROMEDIO"	"(2;88;3;25]"	"0.20477815699658674835004193510205782517 57"
"TASAREPITENCIA"	"[1;4012E5;;08]"	"0.19490445859872630350921332499430449000 7"
"ULTIMOPERIODOCURSADO"	"2004-2"	"0.18980891719745228453997919475708551314 74"
"MATERIASAPROBADAS"	"(26;63]"	"0.14545454545454530769614640680978863564 92"
"MATERIASTOMADAS"	"(30;73]"	"0.13757961783439486855100700448680611886 75"
"NROSEMESTRESCURSADOS"	"(5;11]"	"0.13630573248407633208041556118332826478 21"

"Attribute Name"	"Value"	"Probability"
"PERIODOINGRESO"	"2001-1"	"0.11968680089485458235968034604181817594 71"
"PERIODOINGRESO"	"2001-2"	"0.11744966442953017441149057448407201051 08"
"PERIODOINGRESO"	"2000-2"	"0.11073825503355707078332738991699310521 "
"PERIODOINGRESO"	"2002-1"	"0.10850111856823258837780072611047647484 26"
"ULTIMOPERIODOCURSADO"	"2002-1"	"0.10828025477706995294250624531862943043 64"
"ULTIMOPERIODOCURSADO"	"2001-2"	"0.10828025477706995294250624531862943043 64"
"ULTIMOPERIODOCURSADO"	"2002-2"	"0.10700636942675146573499612368567843570 75"
"PROMEDIO"	"(3;25;3;52]"	"0.10238907849829361063412323254669364005 2"
"PROMEDIO"	"(3;52;3;77]"	"0.09897610921501699346054770881595410833 66"
"PERIODOINGRESO"	"2002-2"	"0.09395973154362402262132756874398589060 4"
"PERIODOINGRESO"	"2003-2"	"0.09172259507829968717977028120487277424 62"
"ULTIMOPERIODOCURSADO"	"2001-1"	"0.09044585987261146593553917533156770478 68"
"PROMEDIO"	"(3;77;4;51]"	"0.08703071672354956148349303584171382356 24"
"ULTIMOPERIODOCURSADO"	"2004-1"	"0.08152866242038213099795663416097374334 12"
"PERIODOINGRESO"	"2004-1"	"0.08053691275167783871914513787371189001 49"
"ULTIMOPERIODOCURSADO"	"2003-2"	"0.07770700636942676583401623367955154284 23"
"ULTIMOPERIODOCURSADO"	"2005-1"	"0.07643312101910826093208171387755121247 3"
"PERIODOINGRESO"	"2003-1"	"0.07494407158836689653583205110415428038 54"
"ULTIMOPERIODOCURSADO"	"2003-1"	"0.06878980891719738677044685836921725559 81"
"ESTRATO"	"4"	"0.06355932203389824174499004532761656573 11"
"ULTIMOPERIODOCURSADO"	"2000-2"	"0.06114649681528665495812672694264809556 31"
"PERIODOINGRESO"	"2000-1"	"0.05704697986577177615007428938831970060 95"
"PERIODOINGRESO"	"20042"	"0.05369127516778519263037486769019600924 59"
"PERIODOINGRESO"	"2005-1"	"0.04474272930648773328646308437961777796 15"
"NROAPOYOSICETEX"	"(4;6]"	"0.04347826086956520394811968833957582255 02"
"MOTIVORETIRO"	"5"	"0.04279749478079326551915763024286881073 12"
"ESTRATO"	"1"	"0.04237288135593217734883572720135065250 88"
"ULTIMOPERIODOCURSADO"	"2000-1"	"0.03057324840764327579245644743948981509 55"
"ESTADOCIVIL"	"Union libre"	"0.03020134228187916647171989699482927121 7"
"PERIODOINGRESO"	"2005-2"	"0.02572706935123041093171981480818778400 93"
"ESTADOCIVIL"	"casado"	"0.02348993288590602019171192473770596661 26"
"PERIODOINGRESO"	"2006-1"	"0.01565995525727068610956654447446319319 25"
"MOTIVORETIRO"	"9"	"0.01043841336116909457741042273667747800 34"
"ESTADOCIVIL"	"separado"	"0.00671140939597314591270086090674438280 4604"
"ESTADOCIVIL"	"error"	"0.00671140939597314591270086090674438280 4604"
"PERIODOINGRESO"	"2004-2"	"0.00559284116331095339599634396118461886 2446"
"ESTRATO"	"5"	"0.00423728813559321342987506193490415299 1595"
"ESTADOCIVIL"	"divorciado"	"0.00335570469798657869264986710255838419 5788"

Tabla 13. Resultados del algoritmo de clasificación por Naive Bayes.

Los resultados anteriores determinan cuál es la probabilidad que existe en la determinación de valor de estudiante inactivo, con relación a todas las variables del modelo y sus respectivos valores. Por ejemplo la variable estado civil, tiene alta afinidad con este resultado (aproximadamente el 93%), dicho en otras palabras esta variable incide en el resultado de estudiante Inactivo en ese porcentaje, 93% son solteros; por contraste el valor de la variable estrato = 5, tiene muy poca incidencia en el estado del estudiante, tan solo se refleja en 0.4%
Las métricas para este resultado son:

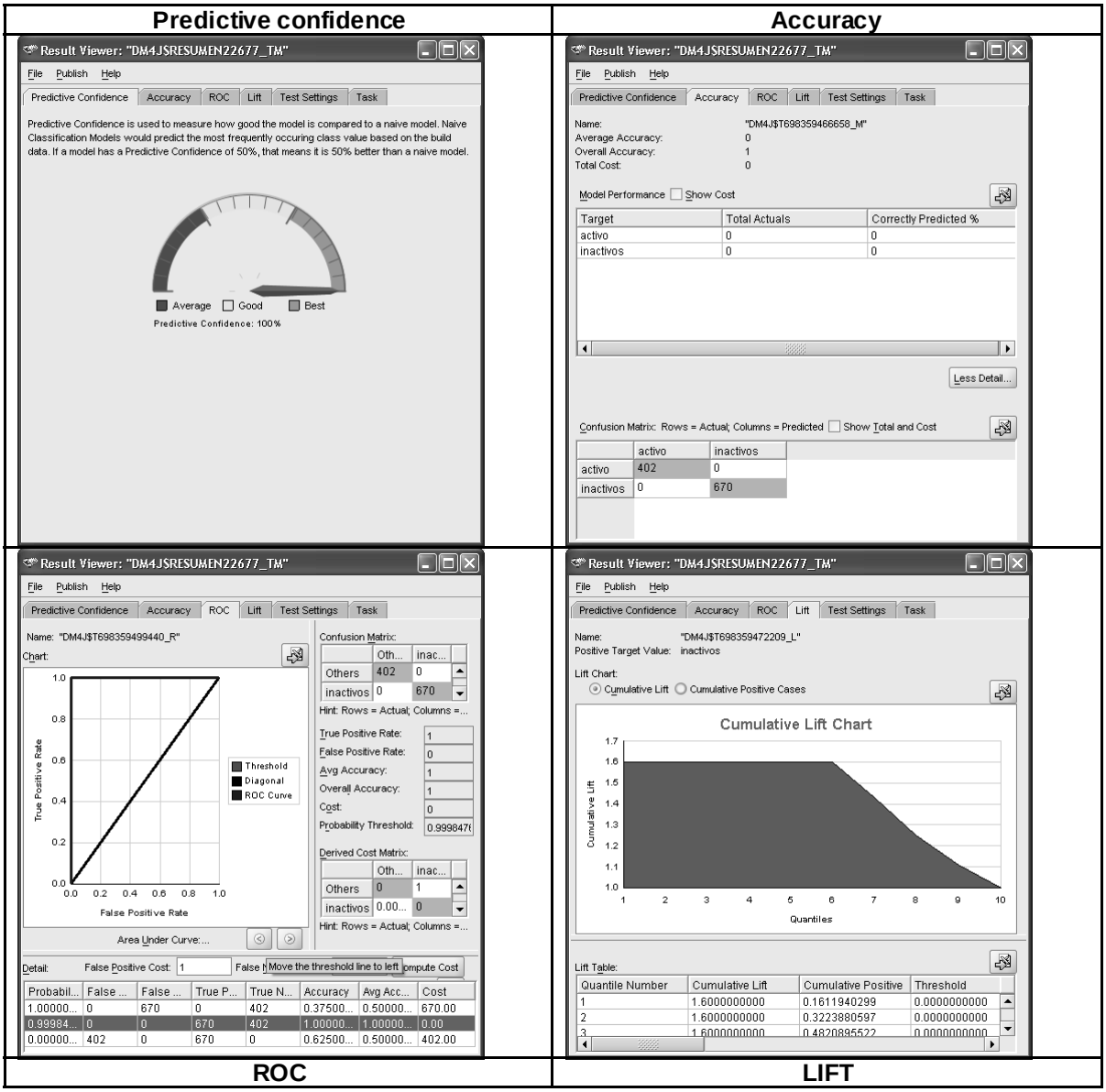

Figura 12. Métricas resultantes de la aplicación del algoritmo de Naive Bayes.

Algoritmo de importancia de los atributos.
Los resultados se muestran a continuación.

"Name"	"Rank"	"importante"
"ULTIMOPERIODOCURSADO"	"1"	"0.753763654"
"MATERIASAPROBADAS"	"2"	"0.205927775"
"MATERIASTOMADAS"	"3"	"0.177271851"
"NROSEMESTRESCURSADOS"	"4"	"0.159186988"
"TASAREPITENCIA"	"5"	"0.116888354"
"PROMEDIO"	"6"	"0.078619883"
"PERIODOINGRESO"	"7"	"0.076528435"
"SEXO"	"8"	"-0.002870417"
"NROAPOYOSICETEX"	"9"	"-0.003030145"
"ESTRATO"	"10"	"-0.003245555"

"ESTADOCIVIL"	"11"	"-0.005182429"
"MOTIVORETIRO"	"12"	"-0.006833474"

Tabla 14. Resultados del algoritmo de importancia de atributos.

La gráfica correspondiente se muestra en la siguiente figura (figura 13).

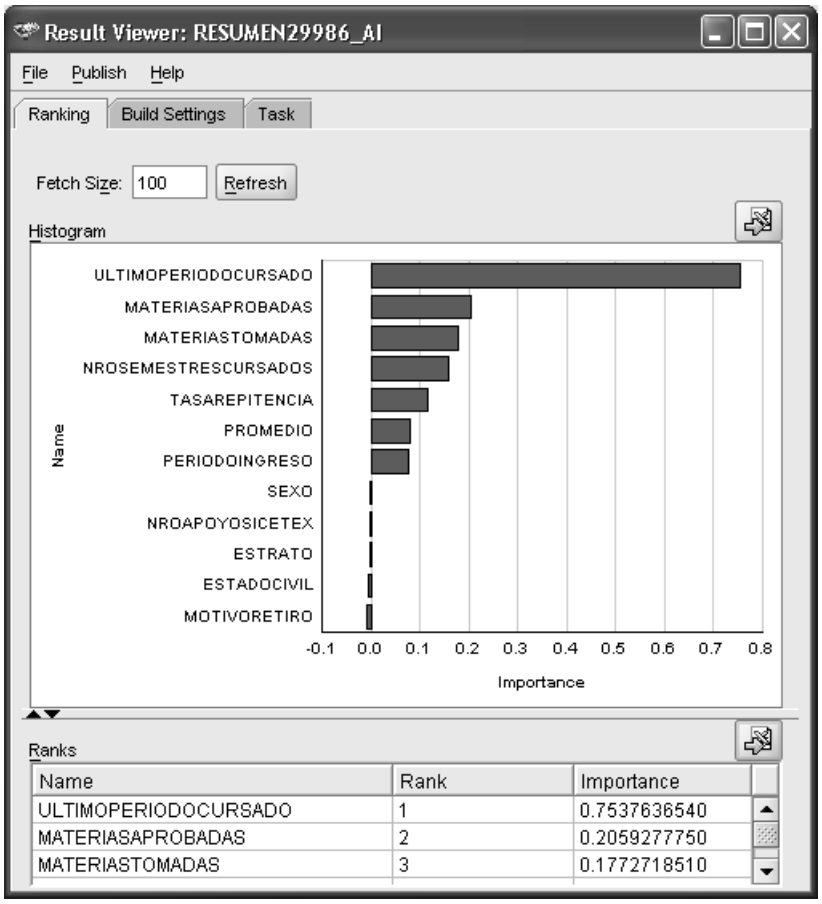

Figura 13. Gráfica resultante de la aplicación del algoritmo de importancia de atributos.

Del análisis de este algoritmo se determinan cuáles son las variables que más inciden en los resultados, en este caso el último período cursado. Pero este resultado es evidente ya que es una de la variables que más completitud tiene en el modelo. Las variables que le siguen en la clasificación sí permiten obtener un mejor reflejo de la incidencia de las mismas y por lo tanto establecer que con solamente las variables materias aprobadas, materias tomadas, tasa de repitencia y promedio puede realizarse el modelo con alta precisión, pero el efecto es particularizar el modelo. Las variables sexo, nroapoyos al Icetex, estrato, estado civil y motivo de retiro no se ven reflejadas significativamente debido a que son variables con alto porcentaje de valores nulos. (variables *scarse).* La decisión en este caso puede orientarse a una de las siguientes dos situaciones: retirar los registros que no tengan valor en estas variables y volver a ejecutar el algoritmo o

buscar la fuente desde donde se pueda tomar los valores faltantes y de todas formas ejecutar de nuevo el algoritmo. En este caso se optó por el de suprimir los valores faltantes.

Algoritmo de extracción de características.
Parámetros:
Tratamiento de desbordamientos:
Puntos de corte: desviación estándar, sigma = 3
Remplazar valores con nulos.
Valores omitidos:
Aplicar a todos los atributos excepto a los escasos.
Para numéricos utilizar la media
Y para categóricos utilizar la moda.

"Attribute Name"	"Value"	"Coefficient"
"ESTADOCIVIL"	"soltero"	"0.462614877928798"
"PROMEDIO"	"null"	"0.374988278271697"
"ESTRATO"	"3"	"0.348662610133059"
"SEXO"	"M "	"0.339148896780529"
"ESTADOESTUDIANTE"	"inactivos"	"0.284006960687346"
"NROSEMESTRESCURSADOS"	"null"	"0.249559531122626"
"TASAREPITENCIA"	"null"	"0.201268857516296"
"ESTADOESTUDIANTE"	"activo"	"0.157557172812243"
"NROAPOYOSICETEX"	"null"	"0.117299498373407"
"ULTIMOPERIODOCURSADO"	"2006-1"	"0.081538805286894"
"MOTIVORETIRO"	"4"	"0.0710351657019039"
"ULTIMOPERIODOCURSADO"	"2004-2"	"0.0708544308662224"
"PERIODOINGRESO"	"2001-2"	"0.0627886528899538"
"MOTIVORETIRO"	"1"	"0.0507339521170848"
"PERIODOINGRESO"	"2002-1"	"0.0337733002442373"
"ESTRATO"	"2"	"0.0308031242086949"
"ULTIMOPERIODOCURSADO"	"2003-1"	"0.0303772356056939"
"PERIODOINGRESO"	"2000-2"	"0.0297705488200751"
"PERIODOINGRESO"	"2003-2"	"0.027490696272961"
"PERIODOINGRESO"	"2005-2"	"0.0268203817414636"
"ULTIMOPERIODOCURSADO"	"2001-1"	"0.0265426458297891"
"ULTIMOPERIODOCURSADO"	"2005-2"	"0.0253066326612151"
"PERIODOINGRESO"	"2002-2"	"0.0248233417140455"
"ULTIMOPERIODOCURSADO"	"2003-2"	"0.0230133254757793"
"MATERIASTOMADAS"	"null"	"0.0221184959791542"
"SEXO"	"F "	"0.0218791736558046"
"PERIODOINGRESO"	"20042"	"0.0217785720640253"
"ULTIMOPERIODOCURSADO"	"2004-1"	"0.0200717246583922"
"PERIODOINGRESO"	"2003-1"	"0.0198335512602865"
"ULTIMOPERIODOCURSADO"	"2002-2"	"0.0197323819323112"
"PERIODOINGRESO"	"2000-1"	"0.0162557490675492"
"ULTIMOPERIODOCURSADO"	"2001-2"	"0.0157971562530138"
"PERIODOINGRESO"	"2005-1"	"0.0153147648088546"
"PERIODOINGRESO"	"2004-1"	"0.014387684968392"
"PERIODOINGRESO"	"2001-1"	"0.0137737933459057"
"PERIODOINGRESO"	"2006-1"	"0.0123152859019535"
"MOTIVORETIRO"	"5"	"0.0118008267890459"
"MATERIASAPROBADAS"	"null"	"0.0111704659350754"
"ULTIMOPERIODOCURSADO"	"2000-2"	"0.0106453424645107"
"ULTIMOPERIODOCURSADO"	"2002-1"	"0.00668205411033281"

"Attribute Name"	"Value"	"Coefficient"
"ULTIMOPERIODOCURSADO"	"2000-1"	"0.00461751147216835"
"ULTIMOPERIODOCURSADO"	"2005-1"	"0.00351383058925184"
"ESTADOCIVIL"	"casado"	"0.00276647086118358"
"PERIODOINGRESO"	"2004-2"	"0.00243288643675976"
"MOTIVORETIRO"	"9"	"0.00134142046509754"
"ESTADOCIVIL"	"Union Libre"	"0.00132503274508156"
"ESTRATO"	"4"	"0.00122277831735348"
"ESTADOCIVIL"	"separado"	"0.00101624279802146"
"ESTRATO"	"1"	"0.000809461988878679"
"ESTRATO"	"5"	"0.000338393425520906"
"ULTIMOPERIODOCURSADO"	"2006-2"	"0.000309978750467603"
"ESTADOCIVIL"	"otro"	"0.000209906376275921"
"ESTRATO"	"6"	"0.0000670641512929997"
"ESTADOCIVIL"	"divorciado"	"0.0000133269648772351"

Tabla 15. Resultados del algoritmo de extracción de características.

La aplicación de los modelos descritos en los párrafos anteriores, (probados y validados), sobre nuevos datos, es la que permite a los usuarios utilizar la potencialidad de la herramienta y encontrar realmente los beneficios que da la minería de datos. El proceso consiste en incluir los nuevos datos (en otra tabla), es decir un nuevo registro de datos con las mismas variables utilizadas por el proceso de creación del algoritmo y seleccionar cual es el algoritmo que se quiere aplicar. Para ello se puede hacer directamente a través de la herramienta utilizando en el menú la opción aplicar (*aply)* y siguiendo al asistente, o por medio de una aplicación Oracle Forms, que interactúa con el modelo y le proporciona los resultados al usuario (desarrollo futuro, como continuación del proceso en la universidad objeto del prototipo).

Los algoritmos a utilizar y la secuencia propuesta es la siguiente:

1. Algoritmo de conglomerados, para obtener en que grupo queda el (los) estudiantes a quienes se les aplica el modelo.
2. Aplicación del algoritmo de clasificación (por árbol o por Naive Bayes) que determina cuál o cuáles estudiantes tienen la condición de ser posibles desertores, es decir quedan clasificados como Inactivos en algún semestre.
3. Para estos estudiantes, se les aplica cada uno de los algoritmos de conglomerados detallados por segmentos de las variables socioeconómicas, personales o laborales, para determinar por comparación y mayor valor, cuál de ellas es la que más influye en el resultado y poder elegir de las políticas disponibles la que dé mayor ayuda al estudiante para evitar que se retire de la universidad.

9.4.4 Sobrevivencia.

Para la obtención de la grafica 14, se tomó el número de estudiantes con base en lapsos de tiempo (cohortes) de 5 años para estudiantes diurnos y 5.5 años para nocturnos. Calculados a partir de 2001-I a 2005- II para estudiantes diurnos y de 2001-I a 2006-I nocturnos. Estas cohortes determinan cómo fue el

comportamiento de la cantidad de estudiantes que ingresaron al comienzo del rango durante el período dado y cómo fue la deserción semestre a semestre. Sumando los valores para un mismo nivel (por ejemplo 1er semestre, 2do semestre, etc) de cada cohorte, se obtiene el valor representativo para cada nivel (semestre) en la universidad.

Luego se calcula el índice de variación de un semestre con relación al anterior tal como lo muestra la tabla 16 y el resultado se aplica para un valor de 100, obteniendo los resultados mostrados en la gráfica de la figura 14.

El resultado para el décimo semestre con relación al primer semestre muestra una deserción del 69%, es decir de 100 estudiantes que entran tan solo 31 terminan.

Semestre	Variación %
1er Semestre	
2do Semestre	14,4
3er Semestre	18,1
4to Semestre	18,5
5to Semestre	10,2
6to Semestre	11,0
7mo Semestre	7,3
8vo Semestre	6,5
9no Semestre	12,4
10mo Semestre	9,4

Tabla 16. Variación porcentual general entre un semestre y el anterior

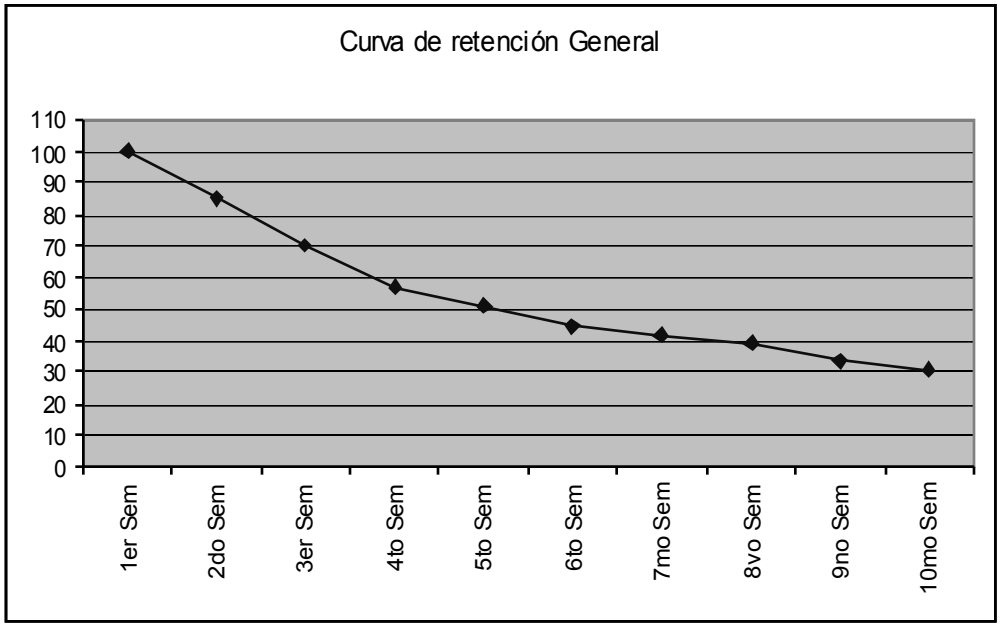

Figura 14. Curva general de retención para un ingreso de 100 estudiantes aplicando los índices de variación mostrados en la tabla 16.

10 Conclusiones.

La aplicación de la minería de datos a una institución universitaria le permite entender en forma particular el comportamiento de sus estudiantes y plantear las estrategias necesarias para controlar y minimizar el problema de deserción. Este entendimiento se logra a través de la aplicación de algoritmos tales como de conglomerados (*clustering*) donde se visualizan los diferentes conjuntos de características similares de estudiantes por orden de importancia; del algoritmo de clasificación por árboles de decisión que por medio de la regla obtenida permite aplicarla a un nuevo estudiante y determinar cuál será el comportamiento de dicho estudiante, es decir dónde quedará clasificado, y el Naive Bayes que permite determinar la probabilidad relacionada de cada una de las variables de estudio con la variable objetivo, por ejemplo el estado del estudiante y clasificarlas por orden de probabilidad, así la probabilidad que el estudiante desertor sea hombre (sexo = 'M') es de 0.72483221476510027101342723968722819697 34 según la tabla 13. De manera similar los algoritmos permiten predecir comportamientos de los estudiantes para aplicar de manera preventiva acciones orientadas a contrarrestar efectos negativos de la situación del estudiante en relación con la universidad.

El estudio planteado en el prototipo se hizo con base en datos históricos existentes en diferentes fuentes de la universidad. Sin embargo otra forma de uso será ver el efecto de las políticas que se apliquen en un determinado momento haciéndole el seguimiento correspondiente. Ésta situación no se ve en éste trabajo ya que no hubo el tiempo necesario para reflejarla en los resultados. De todas maneras con relación a los datos históricos una vez se digiten y lleven al sistema todas las variables, como por ejemplo colegio dónde el estudiante curso su educación secundaria, se podrá confrontar la política de promoción de la universidad con estos datos y determinar si fue o no positiva dicha promoción.

Para el momento en que se realiza este estudio, existe una herramienta que utiliza el Ministerio de Educación Nacional, el SPADIES, la cual permite relacionar a una universidad con relación a las demás. Esta herramienta de ninguna manera contradice o desmiente los valores obtenidos en el presente trabajo, más bien se convierte en una magnifica oportunidad de complemento, pues mientras la primera muestra a la universidad y su entorno, la segunda profundiza en sus características y amplía el entendimiento de su situación. Una característica que no tiene el SPADIES, es la de reflejar los resultados de las políticas que realiza la universidad en su gestión, situación que sí reflejara este estudio, precisamente por la generalidad como lo refleja el SPADIES y el menor numero de variables utilizadas.

Del análisis de los resultados obtenidos, se ve que el comportamiento de cada cohorte de estudiantes no es homogénea, ni en número de estudiantes ni en las características que los identifican, haciendo más complejo su entendimiento.

Para cada uno de los conglomerados definidos por el algoritmo de clasificación es necesario determinar una o unas políticas que refuercen que el estudiante se

quedará (estudiante activo), tal como los del conglomerado 2 definido en la pagina 58 o del conglomerado del grupo 4 (estudiante inactivo) cuya razón de retiro fue la por problemas económicos.

Para efectos de realizar algún tipo de predicción con relación a un estudiante que ingresa, las variables socioeconómicas son las que se deben tener en cuenta ya que se trata de un estudiante que no tiene historial académico en la Universidad. Este modelo al completarse en esta información y tener en cuenta estas variables y las reglas obtenidas de los conglomerados, permitirá incluir al estudiante nuevo en alguno de los grupos de conglomerados definidos por el algoritmo, determinando las políticas a tener en cuenta en cuanto a su desempeño y ayudas.

Si se desea clasificar o hacer algún tipo de predicción con nuevos estudiantes, los modelos a utilizar no deben incluir entre sus variables el periodo de ingreso o el periodo de terminación, ni el número de asignaturas tomadas o el número de asignaturas vistas, etc.

Una facilidad adicional que da la herramienta ODM, en este caso, es el soporte de las métricas de cada uno de los algoritmos, que permite visualizar valores de precisión (*accuracy*), ROC, y LIFT en la obtención del valor de confianza (*confidence*) y predicción.

Como todo proyecto informático el éxito depende de la facilidad de aplicación y de uso por parte de los usuarios. Adicionalmente del compromiso que tengan las directivas de sus instalación y puesta en marcha.

11 Trabajo Futuro.

Ampliar el estudio de similaridad y diferencia con el aplicativo SPADIES.

Aplicar a cada estudiante dependiendo del semestre al que ingrese la probabilidad de continuación o no y determinar con base en ello las políticas necesarias a aplicar para garantizar la continuación de sus estudios en la universidad utilizada en el prototipo.

Ampliar el estudio con otro relacionado con los estudiantes egresados no graduados ya que estos no se ven reflejados en este trabajo y que constituyen otro problema para la universidad.

12 Bibliografía.

[1] Data Mining Applications in Higher Education. Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories. SPSS

[2] La UNPA encabeza un estudio sobre las causas de la deserción educativa. Periódico Austral. Julio 1 de 2005.

[3] El factor educacional como causa potencial de la deserción en primer año de la universidad. Universidad Nacional de Río, Córdoba, Argentina. http://www.unrc.edu.ar/publicar/cde/h21.htm. Artículo bajado de internet el 25 de octubre de 2005.

[4] BOADO, MARCELO. Una aproximación a la deserción estudiantil universitaria en Uruguay. Publicado por Instituto Internacional para la Educación Superior en América Latina y el Caribe. UNESCO.

[5] Latiesa, Margarita. Tipología y causas de la deserción universitaria y el retraso en los estudios.
http://dialogo.ugr.es/anteriores/dial05/11-5.htm. Artículo bajado de Internet el 25 de octubre de 2005.

[6] Estadísticas e indicadores de la universidad Nacional de Colombia, basada en La información estadística del Departamento Nacional de Planeación, Boletín No 27, Educación y fuerza de trabajo, Bogotá, septiembre de 2000, y el Informe de Desarrollo Humano para Colombia, 1999.

[7] Caicedo C., Guarino (2005, Junio 20). COLOMBIA; Alto porcentaje de deserción universitaria; Estadística alcanza el 52% y es considerada como una verdadera 'tragedia nacional'. El Diario La Prensa, p. 14

[8] Jose Solarte. A Proposed Data Mining Methodology and its Application to Industrial Engineering. Agosto 2002

[9] Álvarez Manrique, José Maria. Etiología de un sueño o el abandono de la universidad por parte de los estudiantes por factores no académicos. Tesis de maestría Universidad de los Andes. 1996

[10] Corpoeducación, Sarmiento Gómez, Alfredo; Tovar, Luz Perla; Alam, Carmen. Situación de la educación básica, media y superior en Colombia.
Publicado por: Casa Editorial El Tiempo - Fundación Corona, Fundación Antonio Restrepo Barco. Bogotá, noviembre de 2001.

[11] M. Berry, G. Linoff, “Data mining techniques for marketing, sales and customer relationship management”, Wiley Publishing, Inc. Indianapolis, 2004, pp. 43 – 86.

[12] Hand, David; Mannila, Heikki and Smyth, Padhraic Principles of Data Mining. The MIT Press © 2001 (546 pages)

[13] Malagón Escobar, Luz Miriam, Calderón Cañón, Cesar Augusto y Soto Hernández, Edwin Leonardo. Estudio de la deserción estudiantil de los programas de pregrado de la universidad de los llanos (1998-2004). Villavicencio, Meta Enero de 2006.

[14] Fuente: ministerio de educación nacional SNIES

[15] Universidad Nacional de Colombia. Convenio 107/2002 UN-ICFES. Documento Sobre estado del Arte

[16] Osorio, Ana; Jaramillo, Catalina; Jaramillo, Alberto**.** Deserción estudiantil en los programas de pregrado 1995-1998. Oficina de Planeación Integral. Universidad EAFIT, Medellín, 1999. www.eafit.edu.co/planeacion/final.html

[17]http://www.crisp-dm.org/CRISPWP-0800.pdfy http://www.cs.ualberta.ca/~yli/CRISPDM.ppt

[18] Schumann, Jeffrey A. PhD. Data mining methodologies in educational organizations. UNIVERSITY OF CONNECTICUT. 2005.

[19] Conceptos de Oracle data mining. Manual 10*g* Release 2 (10.2) referencia: B14339-01

[20] Oracle® Data Mining Concepts 10*g* Release 2 (10.2) B14339-01. Junio de 2005.

[21] MontakeFlyerForumB. PDF Oracle Data Mining. Publicado el 17/2/04

[22] D. D. Lee y H. S. Seung, Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* (401, pages 788-791, 1999.

[23] Rodríguez Gama Álvaro. Compilador de las memorias del "Primer Congreso Internacional Sobre Calidad en la Educación, Repitencia, Deserción y Bajo Rendimiento Académico", Un escenario par analizar las causas y proponer soluciones. Bogotá, Septiembre 1 y 2 de 2005

[24] Malagón Escobar Luz Miriam, Calderón Cañón Cesar Augusto, Soto Hernández Edwin Leonardo. Estudio de la deserción estudiantil de los programas de pregado de la Universidad de los Llanos (1998-2004), Universidad de los Llanos. Departamento de Proyección Social, Villavicencio, enero 2006 Universidad de los Llanos. Departamento de Proyección Social

[25] Reyes Ruiz Lizeth. La deserción estudiantil en el programa de psicología de la Corporación Educativa Mayor Del Desarrollo Simón Bolívar, Corporación

Educativa Mayor del Desarrollo Simón Bolívar. Unidad de Autoevaluación del Programa de Psicología. Barranquilla, Colombia, 2000

[26] Girón Cruz Luís Eduardo. González Gómez Daniel Enrique. Determinantes del rendimiento académico y la deserción estudiantil, en el programa de Economía de la Pontificia Universidad Javeriana de Cali. Este artículo es producto de la investigación «Determinantes del rendimiento académico y la deserción estudiantil en el programa de Economía de la Pontificia Universidad Javeriana de Cali», financiada por la Coordinación Institucional de Investigaciones, adscrita a la Vicerrectoría Académica.

[27] Universidad del Rosario. Facultad de Economía. Informe sobre movimiento y deserción estudiantil del pregrado de economía - primer semestre de 2001 a segundo semestre de 2004. Bogotá, 2004

[28] Universidad de los Andes. CEDE. Deserción en las instituciones de educación superior en Colombia.

[29] MINISTERIO DE EDUCACIÓN. Acceder para quedarse: Cobertura con permanencia BOLETÍN INFORMATIVO Nº 6. Enero – marzo, Colombia, 2006

Referencias de Internet:

[30] h]ttp://www.colombiaaprende.edu.co/html/directivos/1598/article-80793.html. Consulta por internet el 15 de agosto de 2006.

[31] Tomado de: http://www.dinero.com/wf_InfoArticulo.aspx?idArt=28020 el 30 de junio de 2006.

[32] ESTUDIO DE LA DESERCIÓN ESTUDIANTIL DE LOS PROGRAMAS DE PREGRADO Universidad del Llano. Tomado de http://www.unillanos.edu.co/ull_insc_web/new_portal/docs/INFORME%20FINAL%20FEBRERO%20DE%202006.pdf el 2 de septiembre de 2006.

Anexos 1. Scripts para la carga de los datos desde las hojas de datos en formato Excel.

Carga de los datos básicos del estudiante.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.DAT_BASICOS
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY "" TRAILING NULLCOLS
(
 IDESTUDIANTE INTEGER EXTERNAL,
 SEXO CHAR,
 FECHANAC DATE,
 DIRECCION CHAR,
 TELEFONO CHAR,
 ESTADOCIVIL CHAR,
 ESTRATO INTEGER EXTERNAL,
 FECHAINGRESO DATE,
 PUNTICFES INTEGER EXTERNAL
)
```

Carga de notas de los estudiante para cada una de las asignaturas cursadas.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.EST_ASIG_NOTA
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY "" TRAILING NULLCOLS
(
 IDESTUDIANTE INTEGER EXTERNAL,
 CODESTUDIANTE INTEGER EXTERNAL,
 PERIODOLECTIVO INTEGER EXTERNAL,
 NOTAASIGNATURA INTEGER EXTERNAL,
 IDASIGNATURA INTEGER EXTERNAL
)
```

Carga la información académica del estudiante.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.EST_SIAC
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY "" TRAILING NULLCOLS
(
 IDESTUDIANTE INTEGER EXTERNAL,
 NOMBREESTUDIANTE CHAR,
 CODESTUDIANTE INTEGER EXTERNAL,
```

```
  PROGRAMA INTEGER EXTERNAL,
  PERIODOINGRESO INTEGER EXTERNAL,
  ESTADO CHAR
)
```

Carga los datos del icfes de los estudiantes.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.DATOSICFES
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY ""
(
  PERIODO CHAR,
  IDESTUDIANTE INTEGER EXTERNAL,
  VALORICFES INTEGER EXTERNAL
)
```

Carga información de la matricula semestre a semestre de cada estudiante.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.MAT_SIAC
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY "" TRAILING
NULLCOLS
(
  IDESTUDIANTE INTEGER EXTERNAL,
  CODESTUDIANTE INTEGER EXTERNAL,
  PERIODOLECTIVO INTEGER EXTERNAL
)
```

Carga información de los prestamos que un estudiante tiene con el icetex.

```
LOAD DATA
APPEND
INTO TABLE RAFAELDM.PRESTAMOSICETEX
FIELDS TERMINATED BY ';' OPTIONALLY ENCLOSED BY ""
(
  IDESTUDIANTE INTEGER EXTERNAL,
  NOMBRE CHAR,
  ESTRATO INTEGER EXTERNAL,
  CANTIDAD INTEGER EXTERNAL
)
```

Cada uno de los procesos anteriores se ejecuto en la ventana de comandos del DOS con la siguiente instrucción estando ubicado el directorio bin del Oracle home.

```
sqlldr userid=RafaelDM/Ora*dm_10-06 control=d:\cargados\Encuestas.ctl
log=d:\cargados\programas.log direct=TRUE data=d:\cargados\Encuestas.csv
```

Anexo 2. Scripts para la depuración de los datos existentes en las tablas una vez cargados los datos.

Para la depuración de los datos se ejecutaron los siguientes scripts.

```
/* actualiza la tabla datosspadies con motivoretiro */

Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  leidos number(4) := 0;
  otros number(6):= 0;
  nom  varchar2(50);
  nomb  varchar2(50);
  est  NUMBER(2):= 0;
  estr  NUMBER(2):= 0;
  id    number(15) := 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;
  CURSOR c1 IS
  select IDESTUDIANTE
  from datosspadies;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
          leidos := leidos + 1;
       if leidos = (leidos/5) * 5 then

         update datosspadies
         set motivoretiro = 1
            where estadoestudiante = 'inactivos' and
             estadoestudiante is null;

         regrabados := regrabados + 1;
       end if;
   exception
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
```

```
/* actualiza la tabla otr con el identificador de datosspadis */

Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  nom  varchar2(50);
  nomb  varchar2(50);
  est  NUMBER(2):= 0;
  estr  NUMBER(2):= 0;
  id    number(15) := 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;
  CURSOR c1 IS
  select IDESTUDIANTE
  from datospadies;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
        update otr
        set idestudiante := c1_rec.idestudiante
         where substr(idestudiante,1,6) = substr(c1_rec.nombres,1,6);

      regrabados := regrabados + 1;
   exception
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;


/* actualiza la tabla de otr con EL idestudiante DE datosspadies */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  nom  varchar2(50);
  nomb  varchar2(50);
```

```
  est  NUMBER(2):= 0;
  estr  NUMBER(2):= 0;
  id    number(15) := 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;
  CURSOR c1 IS
  select IDESTUDIANTE, nombres
  from datospadies;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
        select idestudiante into id
           from otr
        where nombre = c1_rec.nombres;

         update otr
         set idestudiante := c1_rec.idestudiante
          where nombre = c1_rec.nombres;

       regrabados := regrabados + 1;
   exception
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;

/* actualiza la tabla de datosspadies con EL NUMERO DE prestamos icetex */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  nom  varchar2(50);
  nomb  varchar2(50);
  est  NUMBER(2):= 0;
  estr  NUMBER(2):= 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;
  CURSOR c1 IS
```

```
  select IDESTUDIANTE, nombre,estrato, cantidad
  from prestamosicetex;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      select NOMBRES, ESTRATO into nom, est
  from datosspadies
      where idestudiante = c1_rec.idestudiante;
  if est is null then
        estr := c1_rec.estrato;
      else estr := est;
      end if;
  if nom is null then
        nomb := c1_rec.nombre;
      else nomb := nom;
      end if;
        update datosspadies
        set nombres = nomb,
         estrato = estr,
         NROAPOYOSICETEX = c1_rec.cantidad
      where idestudiante = c1_rec.idestudiante;
      regrabados := regrabados + 1;
   exception
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/* actualiza la tabla de datosspadies con EL NUMERO DE SEMESTRES CURSADOS */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  ID  NUMBER(15):= 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;

  CURSOR c1 IS
  select IDESTUDIANTE, count(*) CANT
        from matriculados
        group by idestudiante;
```

```
BEGIN
  FOR c1_rec IN c1 LOOP
   begin

        update datosspadies
        set NROSEMESTRESCURSADOS = c1_rec.cant
      where idestudiante = c1_rec.idestudiante;

      regrabados := regrabados + 1;

   exception

      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
           ||' Otros ==> '|| TO_CHAR(cantidad));
--  COMMIT;
  exception
      when others then
         duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/

/* actualiza la tabla de datosspadies con los datos de estudianteasignaturanota */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  ID  NUMBER(15):= 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  numeroasig number(3) := 0;
  numero_as  number(3) := 0;

  CURSOR c1 IS
  select IDESTUDIANTE, count(IDASIGNATURA) cant_asig
  from estudianteasignaturanota
  where notaasignatura >= 3.00
  group by idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      select idestudiante, MATERIASAPROBADAS,  INTO id, numeroasig
      from datosspadies
      where idestudiante = c1_rec.idestudiante;

         if numeroAsig >= c1_rec.cant_asig then
```

```
            numero_as := numeroasig;
        else numero_as := c1_rec.cant_asig;
        end if;

          update datosspadies
          set materiasaprobadas = numero_as
        where idestudiante = c1_rec.idestudiante;

        regrabados := regrabados + 1;

    exception
        when no_data_found then
         insert into datosspadies (idestudiante, materiasaprobadas)
              values (c1_rec.idestudiante,c1_rec.cant_asig);
          insertados := insertados +1;
        when others then
        cantidad := cantidad + 1;
    end;
   END LOOP;
   DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
          ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
              ||' Otros ==> '|| TO_CHAR(cantidad));
--    COMMIT;
   exception
        when others then
            duplicado := 1;
        DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/

/* actualiza la tabla de datosspadies con los datos de estudianteasignaturanota */
Create or replace procedure actualizaEncuestas is
   cantidad number(6):= 0;
   duplicado number(1) := 0;
   otros number(6):= 0;
   ID  NUMBER(15):= 0;
   insertados number(6) := 0;
   regrabados number(6) := 0;
   numeroasig number(3) := 0;
   numero_as  number(3) := 0;

   CURSOR c1 IS
   select  IDESTUDIANTE,  count(IDASIGNATURA)  cant_asig,  avg(NOTAASIGNATURA)
prom
   from estudianteasignaturanota
   group by idestudiante;
BEGIN
   FOR c1_rec IN c1 LOOP
    begin
        select idestudiante, MATERIASTOMADAS,  INTO id, numeroasig
        from datosspadies
```

```
      where idestudiante = c1_rec.idestudiante;

         if numeroAsig >= c1_rec.cant_asig then
           numero_as := numeroasig;
      else numero_as := c1_rec.cant_asig;
      end if;

         update datosspadies
         set materiastomadas = numero_as,
      promedio = c1_rec.prom,
      where idestudiante = c1_rec.idestudiante;

      regrabados := regrabados + 1;

   exception
      when no_data_found then
       insert into datosspadies (idestudiante, materiastomadas, promedio)
            values (c1_rec.idestudiante,c1_rec.cant_asig, c1_rec.prom);
        insertados := insertados +1;
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/

/* actualiza la tabla de datosspadies con los datos de estudiantes */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  ID  NUMBER(15):= 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  nrosemestres number (2) :=0;
  nrosem      number (2) :=0;
  ultimo_per   varchar2 (8) := 0;
  ultimo      varchar2 (8) := 0;
  perinicial   varchar2 (8) := 0;
  perIngreso   varchar2 (8) := 0;
  CURSOR c1 IS
  select IDESTUDIANTE, NOMBREESTUDIANTE, ESTADOESTUDIANTE,
```

```
        decode( PERIODOINGRESO,20011,'2001-1',20012,'2001-2',20021,'2002-
1',20022,'2002-2', 20031,'
          20032,'2003-2',      20041,'2004-1',20051,'2005-1',      20052,'2005-2',20061,'2006-
1',PERIODOINGRESO) per
  from estudiantes;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      select idestudiante, periodoingreso INTO id, peringreso
      from datosspadies
      where idestudiante = c1_rec.idestudiante;

         if peringreso <= c1_rec.per then
           perinicial := peringreso;
      else perinicial := c1_rec.per;
      end if;

         update datosspadies
         set nombres = c1_rec.nombreestudiante,
      estadoestudiante = c1_rec.estadoestudiante,
      periodoingreso = c1_rec.per
      where idestudiante = c1_rec.idestudiante;

      regrabados := regrabados + 1;

   exception
      when no_data_found then
       insert into datosspadies (idestudiante, nombres, periodoingreso, estado)
            values           (c1_rec.idestudiante,c1_rec.nombreestudiante,           c1_rec.per,
c1_rec.estadoestudiante);
         insertados := insertados +1;
      when others then
      cantidad := cantidad + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
        ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
            ||' Otros ==> '|| TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
      DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/

/* actualiza la tabla de estudianteasignaturanota  */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  referencia number(2) := 10;
```

```
  cant_dup number(5) := 0;
  cant_ot number(5) := 0;
  final  number(2) := 0;
  estado varchar2(15) := null;
  duplicado number(1) := 0;
  CURSOR c1 IS
    SELECT a.IDESTUDIANTE,
    a.codestudiante,
    a.periodolectivo,
    notaasignatura,
    idasignatura
    from est_asig_nota a, estudiantes b
    where a.codestudiante = b.codestudiante
    order by a.IDESTUDIANTE,a.periodolectivo,idasignatura;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      duplicado := 0;
        insert into estudianteasignaturanota
        values(c1_rec.IDESTUDIANTE,
         c1_rec.periodolectivo,
                  c1_rec.idasignatura,
                  c1_rec.notaasignatura);
       if duplicado = 0  then
        cantidad := cantidad + 1;
       else cant_dup := cant_dup + 1;
   end if;
   exception
      when others then
          duplicado := 1;
             cant_dup := cant_dup + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad)    ||'
====' ||' duplicados ==> '|| TO_CHAR(cant_dup));
--  COMMIT;
  exception
      when others then
          duplicado := 1;
          return;
End ActulizadatosSpadies;


/* actualiza la tabla de matriculados  */
Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  referencia number(2) := 10;
  cant_dup number(5) := 0;
  cant_ot number(5) := 0;
  final  number(2) := 0;
  estado varchar2(15) := null;
```

```
  duplicado number(1) := 0;
  CURSOR c1 IS
    SELECT IDESTUDIANTE,
    periodolectivo
    from mat_siac a, estudiantes b
    where a.codestudiante = b.codestudiante
    order by periodolectivo;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      duplicado := 0;
        insert into matriculados
        values(c1_rec.IDESTUDIANTE,
         c1_rec.periodolectivo);
       if duplicado = 0  then
        cantidad := cantidad + 1;
       else cant_dup := cant_dup + 1;
   end if;
   exception
      when others then
          duplicado := 1;
             cant_dup := cant_dup + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad)    ||'
====' ||' duplicados ==> '|| TO_CHAR(cant_dup));
--  COMMIT;
  exception
      when others then
          duplicado := 1;
          return;
End ActulizadatosSpadies;
```

Tambien se creo procedimiento de actualizacion de estudiantes.

```
/* actualiza con la tabla de est_siac  */
Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  referencia number(2) := 10;
  cant_dup number(5) := 0;
  cant_ot number(5) := 0;
  final  number(2) := 0;
  estado varchar2(15) := null;
  duplicado number(1) := 0;
  CURSOR c1 IS
    SELECT IDESTUDIANTE,
    nombreestudiante,
    codestudiante,
    programa,
    periodoingreso,
    estado
```

```
    from est_siac
    where estado = 'inactivos'
    order by idestudiante,periodoingreso;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      duplicado := 0;
        insert into estudiantes
        values(c1_rec.IDESTUDIANTE,
                  c1_rec.nombreestudiante,
               c1_rec.programa,
               c1_rec.estado,
         c1_rec.periodoingreso);
       if duplicado = 0  then
        cantidad := cantidad + 1;
       else cant_dup := cant_dup + 1;
   end if;
   exception
      when others then
          duplicado := 1;
             cant_dup := cant_dup + 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad)    ||'
====' ||' duplicados ==> '|| TO_CHAR(cant_dup));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
          return;
End ActulizadatosSpadies;
/


/* actualiza el la tabla de dat_bas  */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  referencia number(2) := 10;
  cant_dup number(5) := 0;
  cant_ot number(5) := 0;
  final  number(2) := 0;
  estado varchar2(15) := null;
  duplicado number(1) := 0;
  CURSOR c1 IS
    SELECT IDESTUDIANTE,
    SEXO,
    FECHAnAC,
    DIRECCION,
    TELEFONO,
    ESTADOCIVIL,
```

```
      ESTRATO,
      FECHAiNGRESO,
      PUNTiCFES
      from dat_basicos;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
      duplicado := 0;
        insert into dat_bas
        values(c1_rec.IDESTUDIANTE,
        c1_rec.SEXO,
        c1_rec.FECHAnAC,
        c1_rec.DIRECCION,
        c1_rec.TELEFONO,
        c1_rec.ESTADOCIVIL,
        c1_rec.ESTRATO,
        c1_rec.FECHAiNGRESO,
        c1_rec.PUNTiCFES);
       if duplicado = 0  then
        cantidad := cantidad + 1;
       else cant_dup := cant_dup + 1;
   end if;
   exception
      when others then
          duplicado := 1;
   end;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
--   COMMIT;
  exception
      when others then
          duplicado := 1;
          return;
End ActulizadatosSpadies;


/* actualiza el estado del estudiante de la tabla de datosspadies  */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  referencia number(2) := 14;
  final  number(2) := 0;
  estado varchar2(15) := null;
  CURSOR c1 IS
    SELECT    idestudiante,    PERIODOINGRESO,    ULTIMOPERIODOCURSADO,
NROSEMESTRESCURSADOS
    from datosspadies;
BEGIN
  FOR c1_rec IN c1 LOOP
      if c1_rec.ULTIMOPERIODOCURSADO = '2000-1' then
        final := 1;
```

```
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2000-2' then
        final := 2;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2001-1' then
        final := 3;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2001-2' then
        final := 4;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2002-1' then
        final := 5;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2002-2' then
        final := 6;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2003-1' then
        final := 7;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2003-2' then
        final := 8;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2004-1' then
        final := 9;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2004-2' then
        final := 10;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2005-1' then
        final := 11;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2005-2' then
        final := 12;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2006-1' then
        final := 13;
      elsif c1_rec.ULTIMOPERIODOCURSADO = '2006-2' then
        final := 14;
      end if;
      if c1_rec.NROSEMESTRESCURSADOS >= 10 then
        ESTADO := 'activo';
      elsif referencia - final <= 2 then
        ESTADO := 'activo';
      else ESTADO := 'inactivos';
      end if;
  update  datosspadies
        set ESTADOESTUDIANTE = estado
      where idestudiante = c1_rec.idestudiante;
      cantidad := cantidad + 1;
      estado:= null;
      final := 0;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
--   COMMIT;
End ActulizadatosSpadies;
/



/* actualiza la tabla de datosspadies con la informacion existente en matriculadosspdies */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
```

```
  tasa number(4,2) := 0;
  CURSOR c1 IS
    SELECT idestudiante, MATERIASTOMADAS, MATERIASAPROBADAS
    from datosspadies;
BEGIN
  FOR c1_rec IN c1 LOOP
   if nvl(c1_rec.MATERIASTOMADAS,0) > 0 then
      if nvl(c1_rec.MATERIASAPROBADAS,0) = 0
       then tasa := 0;
       else tasa := 1- (c1_rec.MATERIASAPROBADAS / c1_rec.MATERIASTOMADAS);
       end if;
  update  datosspadies
        set tasarepitencia = tasa
      where idestudiante = c1_rec.idestudiante;
      cantidad := cantidad + 1;
   end if;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End ActulizadatosSpadies;
/

/* actualiza la tabla de datosspadies con la informacion existente en matriculadosspdies */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  CURSOR c1 IS
    SELECT idestudiante,Periodo
    from matriculadosspdies
    order by periodo,idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
        update  datosspadies
        set ultimoperiodocursado = c1_rec.periodo
      where idestudiante = c1_rec.idestudiante;
      cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End ActulizadatosSpadies;


/* actualiza la tabla de datosspadies con la informacion existente en matriculadosspdies */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  CURSOR c1 IS
    SELECT idestudiante, Periodo
       from matriculadosspdies;
BEGIN
  FOR c1_rec IN c1 LOOP
```

```
        update datosspadies
        set NROSEMESTRESCURSADOS = nvl(NROSEMESTRESCURSADOS,0) + 1
      where idestudiante = c1_rec.idestudiante;
      cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End ActulizadatosSpadies;

/* actualiza la tabla de datosspadies con la informacion existente en matriculadosspdies */

Create or replace procedure ActulizadatosSpadies is
  cantidad number(4):= 0;
  CURSOR c1 IS
    SELECT idestudiante, MATERIASTOMADAS, MATERIASAPROBADAS
      from matriculadosspdies;
BEGIN
  FOR c1_rec IN c1 LOOP
        update datosspadies
        set MATERIASTOMADAS = NVL(MATERIASTOMADAS,0) +
c1_rec.MATERIASTOMADAS,
          MATERIASAPROBADAS = NVL(MATERIASAPROBADAS,0) +
c1_rec.MATERIASAPROBADAS
      where idestudiante = c1_rec.idestudiante;
      cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End ActulizadatosSpadies;

/* actualiza la tabla de datosspadies con la informacion existente en matriculadosspdies */

Create or replace procedure InsertdatosSpadies is
  cantidad number(4):= 0;
  CURSOR c1 IS
    SELECT idestudiante, PERIODOLECTIVO, FECHANACIMIENTO
      from primiparos;
BEGIN
  FOR c1_rec IN c1 LOOP
        insert into datosspadies (IDESTUDIANTE, PERIODOINGRESO,
FECHANACIMIENTO)
          values (c1.idestudiante, c1.periodolectivo, c1.fechanacimiento)
      cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End InsertdatosSpadies;


/* Actualiza el periodolectivo de los estudiantes segun el periodo en el que entraron */
```

```
Create or replace procedure actualizaEncuestas is
  cantidad number(4):= 0;
  miPeriodolectivo number(6):= 0;
  CURSOR c1 IS
    select idestudiante
    from promedios
     where periodolectivo = 20042
     and nivel = 1;
BEGIN
  FOR c1_rec IN c1 LOOP
     select periodolectivo into miperiodolectivo
     from encuestas
     where idestudiante = c1_rec.idestudiante;
     if miperiodolectivo = 0  then
       update Encuestas
       set periodolectivo = 20042
       where idestudiante = c1_rec.idestudiante;
       cantidad := cantidad + 1;
     end if;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End actualizaEncuestas;
/


Create or replace procedure actualizaEncuestas is
  cantidad number(4):= 0;
  CURSOR c1 IS
    Select idestudiante, count(*)
    from ((select idestudiante,identificacionAsignatura,count(*)
          from asignaturasperdidas
          where rownum < 800
          group by idestudiante,identificacionAsignatura
          having count(identificacionasignatura) > 1)
          order by idestudiante)
    where idestudiante In (
     select distinct idestudiante
     from encuestas)
    group by idestudiante
    order by idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
    update Encuestas
    set numeroasignaturasperdidas = c1_rec.NroAsignaturas
    where idestudiante = c1_rec.idestudiante;
    cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End actualizaEncuestas;
```

```
/
Create or replace procedure actualizaEncuestas is
  cantidad number(4):= 0;
  CURSOR c1 IS
    Select idestudiante,periodolectivo, promedio
    from promedios
    where idestudiante In (
     select distinct idestudiante
     from encuestas)
    group by idestudiante
    order by idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
    update Encuestas
    set promedio = c1_rec.promedio
    where idestudiante = c1_rec.idestudiante
    and ULTIMOPERIODOLECTIVOCURSADO =c1_rec.periodolectivo;
    cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End actualizaEncuestas;
/

/* actualiza el estado del estudiante dependiendo si esta o no el un periodo vigente */
Create or replace procedure actualizaEncuestas is
  cantidad number(5):= 0;
  CURSOR c1 IS
        select count(periodolectivo) cuenta, max(periodolectivo) periodo, idestudiante
        from matriculados
        group by idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
    if c1_rec.ultimo < 20061 then
    update Encuestas
    set Estado = 'Retirado'
    where idestudiante = c1_rec.idestudiante;
    else
    update Encuestas
    set Estado = 'ACTIVO'
    where idestudiante = c1_rec.idestudiante;
    end if;
    cantidad := cantidad + 1;
  END LOOP;
  DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(cantidad));
  COMMIT;
End actualizaEncuestas;
/
```

```
/* actualiza el estado del estudiante dependiendo si esta o no el un periodo vigente */
Create or replace procedure actualizaEncuestas is
  cantidad number(6):= 0;
  duplicado number(1) := 0;
  otros number(6):= 0;
  insertados number(6) := 0;
  regrabados number(6) := 0;
  nrosemestres number (2) :=0;
  nrosem      number (2) :=0;
  ultimo_per   varchar2 (8) := 0;
  ultimo      varchar2 (8) := 0;
  perinicial   varchar2 (8) := 0;
  perIngreso   varchar2 (8) := 0;
  CURSOR c1 IS
  select count(periodolectivo) cuenta,
  min(decode(periodolectivo,20011,'2001-1',20012,'2001-2',20021,'2002-1',20022,'2002-2',
20031,'
        20032,'2003-2',      20041,'2004-1',20051,'2005-1',      20052,'2005-2',20061,'2006-
1',periodolectivo))
   max(decode(periodolectivo,20011,'2001-1',20012,'2001-2',20021,'2002-1',20022,'2002-
2', 20031
        20032,'2003-2',      20041,'2004-1',20051,'2005-1',      20052,'2005-2',20061,'2006-
1',periodolectivo))
  from matriculados
  group by idestudiante;
BEGIN
  FOR c1_rec IN c1 LOOP
   begin
  select                                                    PERIODOINGRESO,
NROSEMESTRESCURSADOS,ULTIMOPERIODOCURSADO
        into perIngreso, nrosem, ultimo
  from datosspadies
  where idestudiante = c1_rec.idestudiante;
      if peringreso = null or
        peringreso = '0' then
        perinicial := c1_rec.per_ingreso;
      elsif peringreso <= c1_rec.per_ingreso then
        perinicial := peringreso;
      else perinicial := c1_rec.per_ingreso;
      end if;
      if nrosem = null or
        nrosem = 0 then
        nrosemestres := c1_rec.cuenta;
      elsif nrosem <= c1_rec.cuenta then
        nrosemestres := cantidad;
      else nrosemestres := nrosem;
      end if;
      if ultimo = null or
```

```
          ultimo = '0' then
          ultimo_per := c1_rec.ult_periodo;
        elsif ultimo <= c1_rec.ult_periodo then
          ultimo_per := c1_rec.ult_periodo;
        else ultimo_per := ultimo;
        end if;
    update datosspadies
    set PERIODOINGRESO = perinicial,
    NROSEMESTRESCURSADOS = nrosemestres,
    ULTIMOPERIODOCURSADO = ultimo_per
        where idestudiante = c1_rec.idestudiante;
        regrabados := regrabados + 1;
     exception
        when no_data_found then
         insert into datosspadies (idestudiante, PERIODOINGRESO,
NROSEMESTRESCURSADOS,
          ULTIMOPERIODOCURSADO) values (c1_rec.idestudiante, c1_rec.per_ingreso,
          c1_rec.cuenta, c1_rec.ult_periodo);
          insertados := insertados +1;
        when others then
        cantidad := cantidad + 1;
     end;
    END LOOP;
    DBMS_OUTPUT.PUT_LINE('Cantidad actualizada ==> ' || TO_CHAR(regrabados)
          ||' ====' ||' Insertados ==> '|| TO_CHAR(insertados)
              ||' Otros ==> '|| TO_CHAR(cantidad));
--    COMMIT;
    exception
        when others then
            duplicado := 1;
        DBMS_OUTPUT.PUT_LINE('Cantidad dup ==> ' || TO_CHAR(duplicado));
End actualizaEncuestas;
/
Create table datosspadies as (
 IDESTUDIANTE                                    NUMBER(15)   NOT NULL constraint
datspadies_PK primary key,
 APELLIDOS                          VARCHAR2(50),
 NOMBRES                            VARCHAR2(50),
 SEXO                              CHAR(2),
 FECHANACIMIENTO                          DATE,
 POSICIONHERMANOS                          VARCHAR2(15),
 VIVIENDAPROPIA                         CHAR(2),
 TRABAJABA                           CHAR(2),
 NIVELEDUCATIVOMADRE                          VARCHAR2(50),
 INGRESOSFAMILIARES                         VARCHAR2(20),
 EDADENICFES                          NUMBER(2),
 NROHERMANOS                            NUMBER(2),
 PUNTAJEICFES                          VARCHAR2(15),
 PERIODOINGRESO                           VARCHAR2(8),
 PROGRAMA                           VARCHAR2(50),
 MATERIASTOMADAS                          NUMBER(3),
```

```
MATERIASAPROBADAS                NUMBER(3),
NROAPOYOSICETEX                  NUMBER(3),
NROAPOYOSFINANCIEROS             NUMBER(3),
NROOTROSAPOYOS                   NUMBER(3),
PERIODOGRADO                     VARCHAR2(8),
PERIODORETIROFORZOSO             VARCHAR2(8),
NROSEMESTRESCURSADOS             NUMBER(3),
TASAREPITENCIA                   FLOAT(5),
ESTADOESTUDIANTE                 VARCHAR2(20))
```