Uso de Minería de Datos Educacional para la determinación de los perfiles de rendimiento académico de los alumnos en la UNNE

Julio C. ACOSTA[1-2], David L. LA RED MARTÍNEZ[1], Carlos A. PRIMORAC[1], Jorge A. GONZÁLEZ[1-2], Mayara F. C. GIMÉNEZ ANTONIOW[2]
(1) Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste
(2) Facultad de Ciencias Agrarias, Universidad Nacional del Nordeste
Corrientes, (3400), Argentina

RESUMEN

Se describe el estado de un Proyecto de Investigación (PI 16F002 acreditado por Res. N° 970/16 CS-UNNE) donde evaluaremos el rendimiento de los estudiantes mediante técnicas de Minería de Datos, para ello analizaremos el perfil de cada estudiante desde otras variables, además de las ya clásicas de: calificaciones y desempeño académico. Describimos el contexto en el que se realiza la experiencia como así también el modelo metodológico propuesto de Matriz de Datos y Sistemas de Matrices de Datos que se adecúa al uso que le damos al Data Warehouse para procesar datos y principalmente determinar las variables que intervienen.

Buscaremos encontrar dichas variables entre otras en: factores socioeconómicos, demográficos, actitudinales; en base a las cuales clasificaremos los diferentes perfiles de alumnos para poder implementar acciones proactivas que contribuyan a mejorar el rendimiento de los alumnos y disminuir la deserción.

Describimos el modelo a implementar con el uso de Data Warehouse para determinar los perfiles de rendimiento académico en las asignaturas Algebra de la carrera Licenciatura en Sistemas de Información (LSI) de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FaCENA) de la Universidad Nacional del Nordeste (UNNE) y Matemática I de la carrera Ingeniería Agronómica (IA) de la Facultad de Ciencias Agrarias (FCA) de la UNNE

Palabras clave: rendimiento académico; almacenes de datos; minería de datos; modelos predictivos.

1. INTRODUCCIÓN

Nuestro proyecto de investigación surge de la necesidad de adoptar acciones proactivas frente al desgranamiento y el bajo rendimiento académico de los alumnos de primer año en la Universidad.

La Universidad y las cátedras en estudio han adoptado diversas medidas tendientes a mejorar los resultados expuestos cuantitativos, tales como un Programa de Tutorías, donde un equipo de docentes y alumnos tutores ejecutan un seguimiento y acompañamiento a los alumnos que se detectan han fracasado en el primer examen parcial y planes de clases de apoyo y consultas extraordinarias en vísperas de parciales y durante el transcurso del dictado de las asignaturas; medidas éstas que no han tenido los impactos deseados.

Las carreras en las que se cursan las asignaturas en cuestión tienen un plan de estudio donde se prevé un régimen de correlatividades, que les pueden generar a los alumnos algunas restricciones para el cursado normal de la carrera; Algebra (LSI) tiene la correlativa Cálculo Diferencial e Integral en el primer cuatrimestre de segundo año y Matemática I (IA) tiene la correlativa Matemática II en el segundo trimestre de primer año.

Ambas asignaturas requieren que la capacidad de razonamiento puro esté involucrada en la enseñanza-aprendizaje, y en el caso de las Matemáticas aplicadas, se procura el conocimiento matemático para usarlos en la aplicación de soluciones concretas y reales de la vida práctica profesional; los alumnos se enfrentan en muchos casos por primera vez, al problema de adquirir conocimiento de modelos matemáticos, para luego aplicarlos en problemas concretos y luego interpretar desde la situación problemática planteada, los resultados obtenidos de los modelos matemáticos usados.

La cantidad de alumnos que regularizan y/o que aprueban las asignaturas involucradas en este proyecto no es satisfactoria, consideramos que esa situación puede contribuir al desgranamiento y deserción de los alumnos en los primeros niveles de sus carreras. Es importante, por tanto, *estudiar y determinar cuáles son las variables que inciden en el rendimiento académico a fin de poder establecer estrategias de acción pedagógicas que permitan mejorar dicho rendimiento*.

El tema de nuestra investigación tiene plena vigencia y actualidad en nuestra Universidad, que tiene políticas definidas de atención y contención a la demanda masiva de parte de los alumnos (principalmente en los primeros años).

El mejoramiento de la calidad académica en la Universidad, no necesariamente debe enfocarse sólo en el sistema de enseñanza-aprendizaje, sino que se debe atender otras variables, como por ejemplo, la sistematización de procesos de evaluación permanentes que permitan monitorear cuestiones ligadas a la calidad académica y retroalimente la propuesta de mejora para la Universidad [1]. El *rendimiento académico* es uno de los factores más críticos que debe evaluarse continuamente.

Definimos rendimiento académico como la productividad del sujeto, matizado por sus actividades, rasgos y la percepción más o menos correcta de los cometidos asignados [2]. Evaluaremos elementos que influyen en el desempeño como: los factores socioeconómicos, la amplitud de programas de estudio, las metodologías de enseñanza, los conocimientos previos del alumno [3]; por esta razón, no resulta adecuado evaluar el desempeño general de los alumnos a través de porcentajes de aprobación, notas obtenidas, etc., ya que esos procesos de evaluación no brindan toda la información necesaria para detectar, y corregir problemas cognitivos, de aprehensión, de discernimiento, actitudinales.

Determinaremos las características propias del estudiante,

analizando patrones de comportamiento y de condiciones que posibiliten la definición de los perfiles de alumnos.

En [4] se presentan varios métodos para determinar y clasificar patrones que se utilizan en Inteligencia Artificial (del inglés Artificial Intelligence - AI) y Aprendizaje de Máquinas (del inglés Machine Learning – ML). La Minería de Datos (del inglés Data Mining - DM), son procesos de descubrimiento de nuevas y significativas relaciones, patrones y tendencias en grandes volúmenes de datos utilizando técnicas de AI y ML. Con estas técnicas se extraen patrones y tendencias para describir, comprender mejor los datos y predecir los comportamientos futuros. En [5] se define DW (del inglés Data Warehouse) como una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar a tomar decisiones. En [6] se expone que la necesidad de proporcionar una fuente única de datos limpia y consistente para propósitos de apoyo para la toma de decisiones y la necesidad de hacerlo sin afectar a los sistemas operacionales son las razones por las que surgen los DW.

Esperamos contribuir a encontrar una respuesta al histórico bajo rendimiento académico de los alumnos. Los modelos predictivos que buscamos, permitirán tomar acciones tendientes a evitar el fracaso académico, detectando los alumnos con perfil de riesgo de fracaso académico de manera temprana, a poco del inicio del cursado de las asignaturas; lo que permitirá concentrar en ellos los esfuerzos de tutorías y apoyos especiales.

En este trabajo se propone la utilización de técnicas de DM, con volúmenes no muy grandes de datos que oscilaran de cientos a miles, sobre información del desempeño de los alumnos de las cátedras Algebra (LSI) FaCENA-UNNE y Matemática I de la FCA-UNNE.

2. MATERIALES Y MÉTODOS

Buscamos detectar grupos de estudiantes en riesgo de fracaso en sus estudios, con la finalidad de tomar medidas proactivas frente al desgranamiento y el bajo rendimiento académico de los alumnos de primer año en la Universidad.

Si bien ambas asignaturas donde se realiza la experiencia tienen régimen de acreditación similar, difieren en la carga horaria y los tiempos de dictado a saber: Algebra (LSI) tiene 128 (ciento veintiocho) horas reloj de dictado de las cuales el 50% corresponde a teoría y el 50 % a trabajos prácticos en la modalidad cuatrimestral (corresponde al primer cuatrimestre de primer año de la carrera), mientras Matemática I (IA) tiene 96 (noventa y seis) horas reloj de dictado con idéntica distribución porcentual de tiempos de dictado de teoría y de trabajos prácticos, pero en la modalidad trimestral (corresponde al primer trimestre de primer año de la carrera).

En ambas asignaturas para alcanzar la condición de alumno regular, los alumnos deben asistir al menos al 75% de las clases de trabajos prácticos, que se dictan dos veces por semana en clases de 2 hs. cada una y deben aprobar 2 (dos) exámenes parciales cuyos contenidos son exclusivamente de trabajos prácticos; cada uno de ellos tiene su instancia de recuperación y para aquellos alumnos que hayan aprobado al menos 1 (uno) de los parciales en cualquiera de las 4 (cuatro) instancias disponibles, existe una instancia más para recuperar el examen que queda aún sin aprobar. Cualquiera de los exámenes parciales se aprueba con 60 (sesenta) puntos sobre 100 (cien) puntos posibles. La asistencia a clases de teoría es libre y se dictan dos veces por semana en clases de 2 hs. cada una.

Se acreditan ambas asignaturas con un examen final al que se accede en condición de alumno regular o de alumno libre; el alumno regular debe rendir en el examen final oral solamente de los contenidos de teoría. El alumno que se presenta al examen final en condición de alumno libre, debe rendir un examen escrito de trabajos prácticos y tras aprobar esa instancia pasa al examen de teoría en condiciones similares a la antes mencionada.

Los porcentuales de los alumnos que regularizan Algebra y Matemática I no son los deseados; en el caso de Algebra, de 320 alumnos inscriptos en los últimos 4 años en promedio, aproximadamente un 20% no alcanza a rendir el primer examen parcial en promedio y al final del cursado, regularizan la asignatura solo un 30% aproximadamente; en el caso de Matemática I el desgranamiento después del primer parcial no es tan evidente y el porcentual aproximado de alumnos regulares al final del cursado es del 40%.

La cantidad de alumnos que regularizan y/o que aprueban las asignaturas involucradas en este proyecto no es satisfactoria, consideramos que esa situación puede contribuir al desgranamiento y deserción de los alumnos en los primeros niveles de sus carreras. Es importante, por tanto, estudiar y determinar cuáles son las variables que inciden en el rendimiento académico a fin de poder establecer estrategias de acción pedagógicas que permitan mejorar dicho rendimiento.

Trabajamos en el desarrollo de métodos que contribuyan a encontrar técnicas para la detección temprana de los alumnos que tendrán dificultades en sus estudios, para ofrecerles contención y acompañamiento especial en el inicio de sus estudios Universitarios. Indagamos en aspectos como:

a) diferencia del nivel de aprendizajes de contenidos previos en los alumnos,
b) situaciones particulares personales de los propios alumnos,
c) la capacidad de las cátedras para el seguimiento del aprendizaje de los alumnos,
d) escasa motivación para el estudio de ciencias básicas y otros que puedan revelarse como incidentes en la problemática que nos ocupa.

Para recuperar contenidos en los grupos de riesgo detectados trabajaremos con materiales elaborados con nuevas tecnologías de la informacion (NTIC). Esto no debe desplazar ni sustituir las formas presenciales de enseñanza - aprendizaje, sino más bien ofrecer alternativas diferentes para aquellos alumnos que requieren modelos diferentes para sus estudios y aprendizajes. En [7] y [8] se considera que las NTIC tienen el potencial para desempeñar un papel importante en la recuperación de contenidos al permitir un abordaje más eficaz, en el sentido de permitirnos procesos de aprendizaje más profundos y más persistentes, mientras el peso de un aprendizaje efectivo permanece con las personas, sus capacidades y valores interpersonales.

Entendemos importante en nuestro trabajo el estudio en dos poblaciones aparentemente diferentes como son los alumnos de las carreras LSI y de IA, para determinar si los perfiles de los

estudiantes varían según la elección de la carrera y medir las diferencias en la predisposición y adaptación para el trabajo y aprendizaje mediado con las NTICs.

En los últimos años se han realizado numerosos trabajos relacionados con la producción de contenidos; en [9] se tiene una concepción global e integral del e-learning, en estos nuevos escenarios se incluyen la combinación del aprendizaje cara a cara y el soportado por medios tecnológicos (especialmente la Web), tal que las fortalezas de ambas configuraciones se puedan aprovechar y explotar. Este aprendizaje combinado (blended learning o b-learning) se considera de suma utilidad no sólo para las universidades sino también para la sociedad en general.

En [10] se ha corroborado que los docentes del siglo XXI deben incorporar definitivamente las NTICs como recursos didácticos, sin abandonar los tradicionales de tiza y pizarrón, pero deben conocer el uso de las NTICs con al menos en parte del potencial que ellas ofrecen; [11] dice que algunas teorías psicológicas y pedagógicas consideran necesaria la inclusión del e-moderator docente con habilidades especiales en las actividades online. La actividad del docente tutor se transforma a veces en un hecho fundamental, dice [12] que la manera en que se usa la tecnología puede transformarse en un factor de gran influencia en la calidad de la EA-EV (enseñanza - aprendizaje en entornos virtuales). Se debe trabajar entonces para lograr una forma de EA-EV que tome en cuenta las necesidades individuales, los intereses y estilos.

En este proyecto de investigación, las variables que inciden en el rendimiento académico de los alumnos serán detectadas a fin de establecer, a través de los valores que ellas toman en cada caso, la población de alumnos en riesgo de fracaso, para establecer acciones tendientes a evitar el fracaso de cada uno de los alumnos, con las acciones que correspondan en cada caso particular y/o de cada grupo detectado y disminuir así el posterior desgranamiento.

3. METODOLOGIA

En [13] identificamos en la información que alimenta nuestra base de datos el modelo de *sistemas de matrices de datos*, según el cual, si asumimos que todo objeto puede ser analizado desde una matriz de datos:

> a) Todos los datos de todas las investigaciones científicas de todas las disciplinas tienen una estructura invariante llamada *matriz de datos* –que se conforma con una *unidad de análisis* (UA), una *variable* (V), una escala de *valores* para las variables (R) y *el indicador*. El indicador aparece en la tesis cuatripartita, [14] niega esta componente, sin embargo introduce lo que él llama *estímulo* (S) y que puede ser identificado como indicador.
>
> b) Todas las investigaciones científicas contienen datos de distinto tipo y de diferentes niveles de integración, existe *un conjunto de matrices de datos* que guardan entre sí relaciones lógico-metodológicas determinadas.
>
> c) El lugar del indicador en la conformación del dato, no es un detalle menor, lo pensamos como aquellos procedimientos aplicados a dimensiones relevantes de la variable para efectuar su medición.

En [13] se identifican elementos de diversos tipos y configuraciones en la descripción de cualquier objeto complejo, y en tal sentido aparece necesariamente un "*grupo de matrices*" formado, por lo menos, por tres matrices de datos: una matriz de datos central, en lo que llamaremos "nivel de anclaje" (N_a), que se focaliza en el plano de la investigación; la unidad de análisis de la matriz de datos del N_a tiene atributos que pueden tratarse a su vez como una nueva matriz de datos, pero ahora en un nivel inferior al N_a, al que llamaremos "nivel sub unitario" (N_{-1}) y una matriz constituida por el contexto de las unidades de análisis del N_a, que denominaremos matriz supraunitaria y se encuentra en un "nivel supra unitario" (N_{+1}).

En [13] se enuncia una *ley general del análisis de datos* según la cual el análisis de datos tiene como tarea invariante *la comparación de un estado de cosas existente (o dado empíricamente) con un estado de cosas posibles en el marco de un modelo (o presunción) asumida como necesaria;* así por ejemplo, cualquier valor estadístico (promedio, frecuencia, mediana, etc.) tendrán sentido solamente si pueden ser *comparados* con algún patrón (conocido o inferido), para estimar hasta qué punto los valores de nuestro estudio coinciden o no con lo esperado y a partir de ello poder estimar la situación presente como contingente o necesaria.

En [13] y [14] se proponen direcciones para el tratamiento y análisis de datos: La dirección del tratamiento y análisis de datos *centrada en la variable*, informa acerca del comportamiento de la población con respecto a alguno de sus aspectos más relevantes; *variable* es lo que se puede predicar de la unidad de análisis y presenta variaciones (de calidad, de orden, de cantidad, de relación) en cada una de las unidades de análisis o de una misma unidad de análisis en diferentes momentos [13]; este tratamiento se hace con procedimientos de estadística descriptiva, va desde el análisis univariado o bivariado hasta los distintos tipos de análisis multivariado [15]; entrega información principalmente sobre la población en estudio, a partir de una muestra y esos valores nos entregarán información de la población, siempre y cuando la muestra sea representativa del universo.

La otra dirección del análisis es *centrada en la unidad de análisis*, la cual nos permite caracterizar los diferentes valores de las variables de cada unidad de análisis, de manera tal que las diferentes configuraciones sean "información" a partir de la cual se pueda inferir una dinámica integral, propia del universo en estudio.

En [13] introduce una tercera dirección de análisis, que denomina *centrado en el valor*; es el análisis en el cual se aplican tratamientos destinados a sistematizar, codificar y/o agregar información, con vistas a la construcción de una variable, la construcción de la variable será un medio más que un fin; ello permite así explicar el tratamiento de la información "desde el origen".

En investigaciones que tratan objetos usuales, esta dirección de análisis puede no evidenciarse, Pero en nuestro trabajo donde no todas las variables que inciden la determinación de los perfiles de rendimiento académico de los alumnos están determinadas, este modo de abordar el tratamiento y análisis de datos tiene una importancia particular y es aquí donde el DW será un importante auxiliar para descubrir cuáles son esas variables.

Tratamiento y análisis de datos centrado en la unidad de análisis y en el valor. No se debe confundir el análisis centrado en el valor con el análisis centrado en la unidad de análisis, porque ambos tienen "sentido horizontal" en la tabla de datos, pero lo cierto es que operan en diferentes niveles, mientras el tratamiento centrado en la unidad de análisis lo hace en el nivel de anclaje, el tratamiento centrado en el valor lo hace en los niveles subunitarios.

El análisis centrado en el valor consiste sintéticamente en:

a) idear criterios para clasificar información cualitativa y exploratoria.
b) ejecutar los procedimientos de resumen que se hayan previsto para sintetizar variables multidimensionales (escalas, índice, tipologías, etc.)
c) reagrupar valores (para poner de manifiesto alguna heterogeneidad respecto de alguna característica relevante).

Esta dirección del análisis responde a tres problemas:

a) la confiabilidad de la información obtenida (de cada medición y del conjunto de las mediciones).
b) la validez de los indicadores elaborados (escalas, índices, tipologías, etc.).
c) el reagrupamiento de valores como efecto de los resultados obtenidos.

Tratamiento y análisis de datos centrado en la dirección de la variable. Esta es la dirección de análisis destinada a sintetizar la información acerca de una(s) variable(s) en particular. Para ello disponemos de las herramientas de la Estadística descriptiva y de la Estadística inferencial. [16] define el campo de aplicación de cada uno de los tratamientos y análisis de datos que se hacen usando métodos estadísticos a saber: la estadística descriptiva entiende en la recolección, ordenamiento, análisis y representación de un conjunto de valores de una variable, con la finalidad de describir las características de la variable; mientras la estadística inferencial, a través de determinados métodos y procedimientos, es capaz de inferir las propiedades de una población o de los elementos de ella a partir del estudio estadístico de una porción de la misma, llamada muestra.

4. DATA WAREHOUSE

Como soporte de los datos trabajaremos con Data Warehouse (DW); en informática, un almacén de datos (DW), es un sistema especial de bases de datos utilizado para el almacenamiento de datos y el procesamiento de los mismos para la presentación de informes y análisis de información, es considerado como un componente central de la inteligencia de organizaciones.

Un DW es un repositorio de datos que proporciona una visión global, común e integrada de los datos, [17] presenta las siguientes características:

a) Orientado a un tema: organiza una colección de información alrededor de un tema central.
b) Integrado: incluye datos de múltiples orígenes y presenta consistencia de datos.
c) Variable con el tiempo: se realizan fotos de los datos basadas en fechas o hechos.
d) No volátil: sólo de lectura para los usuarios finales.

Detrás de la arquitectura de componentes del DW existe un conjunto de procesos básicos asociados: los ETL (del inglés Extract, Transform, Load – Extracción, Transformación y Carga).

Los procesos ETL hacen referencia a la recuperación y transformación de los datos desde las fuentes orígenes cargándolos en el DW. En primer lugar los datos se analizan desde las fuentes y se extraen aquellos que serán de utilidad para el proceso en ejecución.

Luego de extraer los datos se los carga al DW pero, en muchas ocasiones, éstos requieren pasar por un proceso de transformación. La transformación de los datos significa un formateo y/o estandarización de los mismos convirtiendo ciertos números en fechas, eliminando campos nulos, etc.

Es necesario que antes de completar el DW con los datos se realicen controles para enviar información cualitativamente correcta. Luego se procede a aplicar alguna técnica para realizar el análisis de los datos almacenados en el DW. El método más utilizado es el proceso de DM que aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad.

Existen varias alternativas del DM, por ejemplo: la Minería de Datos en Educación (Educational Data Mining, EDM). El objetivo de la EDM es el desarrollo de métodos para la exploración de tipos de datos únicos provenientes de plataformas educativas y usándolos para entender mejor a los estudiantes en el aprendizaje [22]. En [19] [20] [21] y [22] existen diversos estudios y publicaciones que abordan la evaluación de rendimiento académico utilizando técnicas de Minería de Datos.

Modelo propuesto: La estructura del DW, consta de una tabla de hechos y varias tablas de dimensión. Una tabla de hechos o una entidad de hecho es una tabla o entidad que almacena medidas para medir el negocio como las ventas, el coste de las mercancías o las ganancias.

Cada medida se corresponde con una intersección de valores de las dimensiones y generalmente se trata de cantidades numéricas, continuamente evaluadas y aditivas. Se pueden distinguir dos tipos de columnas en una tabla de hechos, columnas de hechos y columnas llaves. Las columnas de hechos almacenan las medidas del negocio que se quieren controlar y las columnas llaves forman parte de la clave de la tabla. Una tabla de dimensiones o entidad de dimensiones es una tabla o entidad que almacena detalles acerca de hechos. Por ejemplo una tabla de dimensión de hora almacena los distintos aspectos del tiempo como el año, trimestre, mes y día. Además incluye información descriptiva sobre los valores numéricos de una tabla de hechos. Las tablas de dimensiones para una aplicación de análisis de mercado, por ejemplo, pueden incluir el tipo de período de tiempo, región comercial y producto. Asimismo las tablas de dimensiones describen los distintos aspectos de un proceso de negocio. Si se desea determinar los objetivos de ventas, se pueden almacenar los atributos de dichos objetivos en una tabla de dimensiones. Cada tabla de dimensiones contiene una clave simple y un conjunto de atributos que describen la dimensión.

En nuestro caso, las columnas de una tabla de dimensiones se utilizan para crear informes o para mostrar resultados de consultas. Por ejemplo las descripciones textuales de un

informe se crean desde las etiquetas de las columnas de una tabla de dimensiones. El modelo que se presenta en este trabajo se compone de la tabla de hechos "ALUMNOS" y varias tablas de dimensiones asociadas a la misma que incluyen características que se desean estudiar.

Etapa de recolección de datos: Tal como se planteó, el estudio del desempeño académico de los estudiantes no sólo debe evaluarse teniendo en cuenta los resultados de las instancias de evaluaciones previstas por la asignatura sino que también deben analizarse otros factores culturales, sociales y/o económicos que afecten el rendimiento del alumno. Por ello para este trabajo resultó determinante la participación directa del estudiante, pues era necesario conocer datos sobre aspectos personales que no se podían obtener de otra manera que no fuera a través de respuestas directas por parte de cada alumno. A tal fin se dispuso la elaboración de una aplicación web que permitió contar con una Encuesta On-Line compuesta por preguntas relacionadas a situación familiar e historial de estudios secundarios, entre otras cuestiones.

Etapa de depuración y preparación de datos: Para la realización de una correcta explotación del DW se debe asegurar que los datos obtenidos en la etapa anterior sean consistentes y mantengan la coherencia entre ellos. Así, en la etapa siguiente, se realizará un proceso de limpieza en los datos, que es la eliminación de aquellos registros con todos sus campos en blanco, corrección de errores tipográficos, llenado de algunos campos nulos, entre otros.

La Encuesta no permite la carga, por parte de los estudiantes, de calificaciones de la asignatura en estudio. Esto se dispuso así para evitar errores en los datos ya sea por olvido, o confusión al momento de ingresar los valores. Por ello la carga de notas correspondientes al primer parcial, segundo parcial y sus recuperatorios, examen final y situación del alumno (regular, promovido o libre), es realizada por el equipo responsable de este trabajo de investigación.

La información se obtendrá a partir de la base de datos histórica de las cátedras continuará respecto a las calificaciones de los alumnos. Con esta información depurada se deberá proceder atrabajar en las próximas etapas: - Carga de Datos al DW: Mediante la ejecución del flujo de datos, la información almacenada en la tabla *encuesta* se distribuirá a las tablas pertenecientes al modelo del DW.

4. RESULTADOS

Hasta el momento se ha completado la primera etapa que implicó el diseño del modelo del DW sobre el cual se implementarán técnicas de DM a fin de encontrar las principales variables que intervienen en el rendimiento académico de los alumnos para así determinar los perfiles de rendimiento académico de los estudiantes, vinculados a su desempeño académico en las asignaturas LSI-FaCENA e IA-FCA UNNE. En el avance que aquí se presenta respecto del Proyecto se pudo comprobar que la etapa de depuración y preparación de los datos ha demandado tiempo y esfuerzo debido principalmente a la poca integridad y coherencia que existía en la información recolectada que luego se usará para realizar la evaluación final. En etapas sucesivas se continuará con el proceso de minería de datos para evaluar y comparar patrones que se obtengan, incorporando eventualmente nuevas variables detectadas para definir los perfiles de estudiantes. La evaluación, análisis y utilidad de estos patrones con los que se construirá un modelo predictivo de rendimiento académico permitirá soportar la toma de decisiones eficaces por parte del cuerpo docente de las asignaturas involucradas.

Agradecimientos: Este trabajo es sostenido por el Proyecto de Investigación "Incidencia de los perfiles de los alumnos en el rendimiento académico en Matemática del primer año de la Universidad", código 16F002 de la Universidad Nacional del Nordeste (Argentina).

6. REFERENCIAS

[1] Briand, J. Daly, J. Wüst, "A unified framework for coupling measurement in objectoriented systems". **IEEE Transactions on Software Engineering**. Vol. 25 (1), pp. 91-121, 1999.

[2] J. Maletic, M. Collard, A. Marcus, "Source Code Files as Structured Documents". **Proceedings 10th IEEE International Workshop on Program Comprehension (IWPC'02)**, pp 289-292, París, 2002.

[3] A. Marcus, **Semantic Driven Program Analysis**. Kent State University Doctoral Thesis OH, USA, 2003.

[4] A. Marcus, J. Maletic, "Recovering Documentation-to-Source-Code Traceability Links using Latent Semantic Indexing". **Proceedings 25th IEEE/ACM International Conference on Software Engineering (ICSE'03),** Vol. 3(10), pp. 125-137, USA, 2003.

[5] G. Salton, **Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer**. Addison-Wesley Longhman Publishing Co., Boston, 1989.

[6] J. Molina López, J. García Herrero, **Técnicas de Análisis de Datos**. Universidad Carlos III, Madrid, 2006.

[7] R. Motsching-Pitrik, A. Holzinger, "Student-centered teaching meets new media: concept and case study". **Journal of Educational Technoplogy and Society,** Vol. 5(4), pp. 160-172, 2002.

[8] M. Derntl, T. Hampel, R. Motschnig-Pitrik, T. Pitner, "Inclusive social tagging and its support in Web 2.0 services". **Computers in Hum an Behavior,** Vol 27(4), pp. 1460-1466, 2011.

[9] M. Nichols, "A theory for e-Learning". **Journal of Educational Technology and Society**, Vol (2), pp. 1-10, 2003.

[10] J. Acosta, D. La Red Martínez, **Un aula virtual no convencional de Algebra en la FaCENA-UNNE**. Ed. Académica Española, Saarbrücken, Germany, 2012.

[11] G. Salmon, **E-moderating: The key to teaching and learning online.** Kogan Page, London, 2000.

[12] E. Wenger, D. White, J. Smith, **Digital habitats. Stewarding technology for communities**. Cpsquare, Portland, USA, 2009.

[13] J. Samaja, **Epistemología y Metodología: elementos para una teoría de la investigación científica**. Eudeba, Buenos Aires, Argentina, 2005.

[14] J. Galtung, **Teoría y método de la investigación social. Tomo I (5ta ed.)**. Eudeba, Buenos Aires, Argentina,1978.

[15] R. Ynoub, R. **El proyecto y la metodología de la investigación científica.** Cengage Learning, Buenos Aires, Argentina, 2007.

[16] R. Johnson, P. Kuby, **Estadística Elemental. Lo esencial.** International Thomson Editores, México DF, 2003.

[17] J. Curto Dias, **Introducción al business intelligence.** UOC, Barcelona, España, 2010.

[18] R. Baker, K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions". **Journal of Educational Data Mining,** Vol 1, pp. 3-16, 2009.

[19] S. Formia, L. Lanzarini, "Caracterización de la deserción universitaria en la UNRN utilizando minería de datos". **Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología (TE&ET)**, Vol. (11), pp. 92–98, 2013.

[20] R. Pereira, A. Romero, J. Toledo, "Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos". **Vínculos,** Vol. (10) 1, pp. 374-383. Universidad Distrital Francisco José de Caldas, Colombia, 2013.

[21] D. La Red Martínez, J. Acosta, V. Uribe, A. Rambo, "Academic Performance: An Approach From Data Mining". **Journal of Systemics, Cybernetics and Informatics**, Vol. 10 N° 1 pp. 66-72, U.S.A, 2012.

[22] D. La Red Martínez, M. Giovannini, M. Báez Molinas, J. Torre, N. Yaccuzzi, "Academic performance problems: A predictive data mining-based model". **Academia Journal of Educational Research**, Vol. 5, 4 pp. 61-75. England, U.K. 2017.