Business Intelligence applied to *Learning Analytics* in student-centered learning processes.

Guido Riofrio, Eduardo Encalada, Daniel Guamán
Universidad Técnica Particular de Loja,
Loja, Ecuador
geriofrio@utpl.edu.ec, aeencalada@utpl.edu.ec,
daguaman@utpl.edu.ec

Jose Aguilar
Universidad de Los Andes, Mérida, Venezuela.
Prometeo researcher, Universidad Técnica Particular de Loja, Ecuador.
aguilar@ula.ve

Abstract .- This work aims to evaluate the use of Learning Analitycs (LA) in Higher Education; students, and attributes such as: profile, interactions in a virtual learning environment learning - for this use business intelligence paradigm in order to explore and exploit the data from one of the actors in the educational process is analyzed , test scores, among others, which will contribute to their educational success. In particular, this paper tries to answer the following specific objectives: To identify factors that influence the decision of a college student distance learning to abandon their studies and get the profile of potentially susceptible students from their university studies. To meet this purpose we define two analysis tasks learning and use a business intelligence methodology to implement it.

Keywords—Business Intelligence; Learning Analytics; Online Learning, Data Mining

I. INTRODUCCION

La Inteligencia de Negocio (BI, Business Intelligence, por sus siglas en inglés) tiene un potencial uso en las organizaciones que poseen mucha información y conocimiento oculto, lo cual les permite alcanzar objetivos estratégicos y tácticos [1, 3, 8]. Pero este paradigma debe ser utilizado correctamente, basado en los objetivos, procesos y situaciones adecuadas de la organización, lo que permita producir información y conocimiento requerido para la toma de decisiones y de esta forma identificar tendencias, oportunidades, riesgos, entre otros. Uno de los aspectos claves en un proyecto de BI es determinar la situación actual y objetivo, así como indicadores que puedan ser utilizados en el análisis estratégico.

BI involucra la gestión de datos, información y conocimiento de la organización, así como el modelado de sus procesos. En particular, los datos deben ser extraidos, integrados, depurados, eliminar inconsistencias, etc., pero uno de los aspectos fundamentales para generar conocimiento a partir de los datos es utilizar tareas de minería de datos y OLAP (On-Line Analytical Processing, por sus siglas en inglés).

Actualmente existe un amplio uso de BI en diferentes ámbitos (industrial, comercial, económico, financiero, etc.), gran parte relacionados con la tecnología de la información (herramientas de software y plataformas). Ahora bien, es poco lo que se ha hecho dentro del contexto educativo, por lo tanto, en este dominio es importante analizar su uso en tareas de LA.

Se puede definir a LA como la recopilación, análisis, uso y visualización de los datos que son parte de un contexto educativo, a partir de los cuales se podrá construir modelos que permitan analizar y sugerir la mejora del proceso enseñanza - aprendizaje [2]. En este trabajo se propone investigar el uso de BI en tareas de LA enfocados en el estudiante y su aprendizaje en línea. Básicamente, BI le provee a LA los datos desde diversas fuentes; con ellos se realiza actividades como modelado de datos, almacenamiento de datos, seguridad de datos, etc. por ende se requiere utilizar un correcto proceso de extracción, transformación y carga (ETL, Extraction, Transformation and Load, por sus siglas en inglés) con los datos de la organización educativa.

En este artículo vamos a describir el proceso de BI vinculado al dominio de LA, lo cual involucra los siguientes aspectos: la caracterización de las situaciones objetivo, la especificación de indicadores de LA, la definición del modelo de datos, el tratamiento de los datos para construir la vista de datos sobre los cuales se aplicará el proceso de LA, y la definición de las operaciones OLAP y tareas de minería de datos que responden a los indicadores del proyecto de LA del presente estudio.

Esta especificación de BI para LA será usada en un caso de estudio que corresponde al proceso de aprendizaje en línea centrado en el estudiante de una modalidad de estudios a distancia, el mismo que como instrumento de apoyo utiliza un Entorno Virtual de Aprendizaje (VLE, Virtual Learning Environment, por sus siglas en inglés). En dicho caso de estudio se propone como objetivos del proyecto de LA: i) Identificar los factores que influyen en la decisión de un estudiante de abandonar sus estudios superiores, y ii) Obtener el perfil del alumno potencialmente propenso a abandonar sus estudios.

El resto del trabajo se organiza de la siguiente manera: la sección 2 presenta los aspectos teóricos claves de nuestra propuesta vinculados a BI y LA. La sección 3 caracteriza el proceso de BI en LA para el caso de estudio, la sección 4 se dedica a analizar los objetivos derivados del proceso de LA obtenidos por el proceso de BI, y la sección 5 discute y analiza las conclusiones.

978-1-4673-9143-6/15/$31.00

II. BASE TEORICA

A. Inteligencia de Negocio

Un elemento fundamental en todo proceso de planificación estratégica organizacional, consiste en definir metas u objetivos del negocio. Nosotros proponemos modelar cada objetivo del negocio como una situación objetivo. Intuitivamente, una situación objetivo define un estado del mundo deseado en términos de atributos, sus propiedades, y las interrelaciones entre ellos.

BI tiene como finalidad analizar los datos e información del negocio con el objetivo de apoyar y mejorar los procesos de toma de decisiones a nivel organizacional, en particular, otorgando el conocimiento de forma visual para ayudar en la decisión. BI explota todos los datos de la organización, por ejemplo, los sistemas ERP. BI tiene el potencial de maximizar el uso de la información y el conocimiento para crear una plataforma inteligente de análisis de la organización [1, 3, 8].

Normalmente, BI recupera los datos de los procesos operativos diarios, y los transforma en información y conocimiento [3, 8]. Las principales características de un proyecto de BI son: la capacidad de proporcionar información representativa para los altos niveles gerenciales de una organización, para apoyar sus actividades estratégicas (fijar metas, prevenir, planificar, etc.), y de seguimiento (analizar e integrar datos, etc.). Para ello, BI usa datos en tiempo real como históricos de los diferentes componentes organizacionales (por ejemplo, sistemas ERP). Algunas características de BI son:

- Se basan en la extracción, análisis y visualización de la información.
- Utiliza herramientas capaces de recoger, procesar, almacenar y recuperar datos de diferentes fuentes, interfaces gráficas, técnicas OLAP y de minería de datos, etc.
- Se orienta a buscar oportunidades del negocio, para la toma de decisiones estratégicas.

En [1, 7] se describe de manera detallada los componentes y arquitecturas de BI. En general, un ciclo de vida de BI sigue el desarrollo tradicional de un proyecto de software. Ahora bien, un aspecto fundamental que lo diferencia son los datos, que pueden ser explotadas por diferentes estudios de BI si se extraen, procesan y almacenan correctamente, y se mejoran en versiones sucesivas.

B. Análisis de aprendizaje (LA)

LA puede ser definida como el uso de los datos producidos por los estudiantes durante sus procesos de aprendizaje, con el fin de construir modelos, patrones, etc., que permitan analizar y mejorar dichos procesos mediante la predicción y el asesoramiento a los estudiantes, el descubrimiento de conocimientos e información necesaria durante su aprendizaje, entre otras. Son múltiples los intereses de usar LA [2, 9]:

- Para las organizaciones educativas, para mejorar sus cursos actuales, ofertas curriculares;
- Para los administradores de las instituciones educativas, para tomar decisiones de mercadeo, contrataciones, evaluación de rendimiento, entre otras.
- Para los estudiantes, para mejorar sus procesos de aprendizaje, sus patrones de aprendizaje, logros, entre otros.
- Para los profesores, para identificar a los alumnos en situación de riesgo (que pueden abandonar o fracasar), entre otros.

LA [2] combina datos institucionales, modelos (predictivos y descriptivos, etc.), técnicas de análisis estadístico, entre otras, para crear conocimiento para los estudiantes, profesores o administradores de organizaciones educativas, con el fin de mejorar el desempeño académico. Algunos aspectos que aumentan el interés en LA son:

- El importante número de tipos de plataformas educativas en línea: VLE, sistemas de gestión académica, etc.
- La cantidad importante de datos que estas plataformas recogen sobre los estudiantes, el proceso de aprendizaje (por ejemplo, VLE), etc., interesante para aplicar técnicas de BI sobre ellos.
- La necesidad de disponer de evidencias sobre el progreso del proceso enseñanza - aprendizaje en tiempo real.
- El deseo de una organización educativa de mejorar su modelo educativo a fin de brindar una educación de alta calidad.

Algunas áreas de investigación muy importantes para LA son [2, 9]: minería de datos, análisis web, análisis de redes sociales, inteligencia artificial y estadística. Algunas herramientas de software de LA son [2, 9]: SNAPP, LOCO-Analyst y BEESTAR INSIGHT.

C. Metodología para procesos de BI

En [8] se ha propuesto una metodología para proyectos de BI la cual considera definir las situaciones objetivo, los indicadores para saber si estas se alcanzan, y las tareas de minería de datos/semántica para calcular esos indicadores. Normalmente, las situaciones objetivos describen estados específicos de la organización que mejoran sus actuaciones. Por lo general, con los datos operacionales es imposible obtener respuestas para analizar y tomar decisiones que posibiliten alcanzar dichas situaciones. Se requiere de información estratégica que debe ser extraída a partir de los datos operativos.

Una organización exitosa analiza estas situaciones estratégicas, y define indicadores con el fin de ayudar a la organización a tomar las decisiones correctas para llegar a estas situaciones. En general, estos indicadores cuantifican diversos aspectos de las actividades y dinámicas de la organización, por lo que es necesario monitorear permanentemente la organización. Los indicadores en un BI normalmente se obtienen a partir de tareas de minería de datos/semánticas o de operaciones OLAP. Basado en estas ideas, en [2] se propone una metodología para el desarrollo de proyectos de BI que está compuesta por los siguientes pasos [8]:

Etapa 1: Definición de las situaciones de destino

En este caso, definimos las principales preguntas que el proyecto de BI debe responder. Normalmente, las situaciones objetivo son las preguntas estratégicas que la organización debe responder. Estas situaciones objetivo justifican el proyecto de BI, y definen las necesidades y oportunidades de negocio identificadas. Normalmente, se requiere la extracción de un conocimiento oculto en los datos operacionales de la organización, mediante tareas de minería de datos/semánticas. Ellos son muy importantes, porque definen los indicadores que se buscan obtener. El objetivo principal en esta fase es definir las situaciones objetivos. Estas situaciones describen diferentes tipos de estados en una organización: fortalezas, debilidades, oportunidades y amenazas (FODA).

Para analizar estos estados, y en particular para saber si la organización está cerca de ellos, normalmente, se definen indicadores. Más específicamente, se identifican indicadores de lo que puede ser observable. La definición de indicadores se basa en la misión de la organización, y por supuesto, en las situaciones objetivos a alcanzar, y requiere el descubrimiento de conocimiento estratégico. El análisis de esos indicadores es una tarea importante en cualquier proceso de planificación estratégica. Durante el proceso de planificación estratégica, los indicadores permiten la toma de decisiones en los puntos de decisión. En cada punto, una opción se elige entre un grupo de opciones disponibles utilizando la información y el conocimiento proporcionan por los indicadores.

Etapa 2. Modelo de datos del proyecto de BI

Ahora se deben preparar los datos a utilizar para calcular los indicadores. Normalmente, durante esta etapa se diseña el almacén de datos que contiene tanto los datos históricos y actuales, optimizados para hacer consultas y análisis rápido. Los almacenes de datos se basan en el diseño de modelos de datos multidimensionales, con el fin de tener en cuenta los principales datos necesarios para responder a las situaciones objetivos. Sistemas de bases de datos relacionales tradicionales pueden manejar estas situaciones, pero usando varias consultas (en muchos casos, las consultas son tan complejas que son difíciles de mantener). Los almacenes de datos requieren la extracción, transformación y procesamiento de datos para una integración y análisis de alto nivel, por estas razones, en esta etapa se definen estas tareas. El almacén de datos prepara aquellos que se utilizarán en el proyecto de BI (los limpia, corrige, normaliza, etc.). Algunos de los pasos en esta etapa son:

- Analizar los datos operativos de la organización, es decir, en este paso se identifican las fuentes de datos, con sus diagramas ER, atributos y referencias entre datos.
- Diseñar la base de datos multidimensional, que define el modelo lógico y físico de datos. El modelo de datos se utiliza para definir el almacén de datos
- Diseñar el proceso de ETL, que se utilizará para recuperar los datos desde las diversas fuentes de datos.
- Ejecutar el proceso ETL sobre los datos operacionales de la organización, para concebir la vista minable operativa sobre la que trabajará la siguiente etapa.

Los modelos multidimensionales representan una extensión del modelo relacional y, normalmente, se basan en un esquema en estrella. Consiste en establecer la relación entre dimensiones que describen objetos de información de interés, a través de una tabla de hechos que contiene las medidas a obtener. Esto define un modelo de cubo basado en n dimensiones que utiliza una visión multidimensional sobre un conjunto de datos individuales. El modelo multidimensional permite calcular los indicadores de rendimiento.

Etapa 3. La extracción de conocimiento (indicadores)

Sobre la base de las situaciones objetivos del proyecto de BI, en esta etapa se definen diferentes tipos de modelos o métricas para la interpretación estadística y el análisis de las situaciones objetivos, utilizando los indicadores respectivos. En este nivel se utilizan tecnologías como OLAP, minería de datos, etc. El motor OLAP es un generador de consultas para explorar y analizar información detallada resumida en bases de datos multidimensionales. Las operaciones típicas OLAP son: enrollar (roll-up o drill-up, por sus siglas en inglés), profundizar (drill-down o roll-down, en inglés), o hacer picadillo (slice and dice, en inglés) (para más detalles sobre estas operaciones, ver [1, 3]). OLAP proporcionan herramientas de análisis para encontrar tendencias dentro de los datos, pero no descubre relaciones ocultas o patrones. Para estas tareas se requieren herramientas más potentes, como las técnicas de minería semántica/datos. Los patrones o modelos descubiertos por ellos se validan y luego se convierten en herramientas operacionales para ser utilizados en los procesos de toma de decisión. Las herramientas OLAM (on-line de minería de datos analítica) son sistemas OLAP utilizados para minería de datos en datos multidimensionales.

La definición de las operaciones OLAP es muy compleja, porque deben dar un nuevo conocimiento oculto en el almacén de datos. Debe analizarse qué datos se deben cortar y rebanar, o que datos se deben girar o rodar, a lo largo de los niveles jerárquicos. Así, es necesario un modelo multidimensional en el que estas operaciones se pueden realizar fácilmente, en tiempo real, para los calcular indicadores estratégicos.

En cuanto a la minería de datos o semánticas, son aún más complejas. Existen metodologías específicas para este paso. El principal aspecto a considerar es que las tareas de minería de datos/semánticas deben estar muy bien definidas, y para cada uno hay un paso muy importante, la preparación de los datos , la cual esta compuesta por dos aspectos: la definición de la visión conceptual de los datos (los diferentes atributos que se deben considerar para construir los modelos descriptivos y predictivos), y la construcción de la vista operativa de los datos (es decir, la extracción de los datos para construir los modelos). Si estas vistas son incorrectas, los resultados de la minería de datos/semánticas serán también incorrectos.

III. APLICACION DE LA BI EN LA PARA PROCESOS DE APRENDIZAJE EN LINEA CENTRADA EN ESTUDIANTES

A continuación vamos a aplicar la metodología de la sección II.C. en tareas de LA para un proceso de aprendizaje en línea centrado en el estudiante.

A. Etapa 1: Definición de las situaciones de destino

El proceso de aprendizaje en línea es el proceso de formación en un entorno virtual donde los profesores y los estudiantes están en diferentes espacios, y tal vez tiempos [4, 5, 6]. Este proceso puede dirigirse y adaptarse a las características de los diferentes estudiantes (centrado en el estudiante), y cuenta con diferentes interfaces y metodologías de aprendizaje en línea específica. Algunos entornos virtuales en línea por lo general implican sistemas de gestión del aprendizaje (LMS por sus siglas en inglés Learning Management Systems, VLE por sus siglas en inglés Virtual Learning Environment como Moodle, etc. La situación objetivo es:

*Determinar los factores influyentes en la deserción estudiantil en este tipo de procesos de aprendizaj*e.

Al respecto, algunas características a considerar de cada estudiante son:

- Titulación
- Centro de estudios
- Género
- Edad
- Zona geográfica (Cantón, Provincia)
- Tipo de discapacidad
- Tamaño del centro de estudios (volumen de estudiantes)
- % Avance en el programa formativo

Otros aspectos a considerar es el umbral de avance a nivel de la carrera bajo el cual se produce el mayor porcentaje de deserción. Basado en ello, algunas preguntas importantes de este proceso son:

- ¿Cómo influye el rendimiento académico del estudiante en la deserción?
- ¿Existen titulaciones cuyos índices de deserción implican decisiones urgentes?
- ¿Es posible predecir cuando un estudiante es propenso a abandonar sus estudios?
- ¿La ubicación geográfica del estudiante es un factor determinante en la deserción?

Desde esas preguntas se derivan los indicadores que nos permiten caracterizar la situación objetivo, es decir establecer los factores influyentes en la deserción. Estos son:

- Índice de deserción por carreras
- Centros de estudio con mayor índice de deserción
- Tasa de efectividad de aprobación (hitos alcanzados versus número de intentos)
- Índice acumulado de deserción según nivel de avance
- Índice de deserción según rango de edad de estudiantes

B. Etapa 2. Modelo de datos del proyecto de BI

Lo primero que definimos es el modelo multidimensional dimensional (vista conceptual) y después la vista operativa, para lo cual desarrollamos las operaciones ETL desde las fuentes de datos de la organización educativa.

Modelo multidimensional (esquema estrella)

La figura 1 presenta el diseño dimensional para nuestro caso de estudio, el cual se centra en el histórico de los estudios cursados (hechos) por cada estudiante, caracterizando cada estudio cursado en función de varios factores (dimensiones). De esta manera, partimos de un modelo tipo estrella en el que la tabla de hecho contiene los indicadores a calcular para la situación objetivo, y los apuntadores a las diferentes tablas de dimensiones requeridas, las cuales son:

- Estudiante: información básica de los estudiantes que cursan estudios en una o más titulaciones, y a quien pertenece el expediente caracterizado en ese momento por la tabla hecho.
- Titulación: contiene la información que describe las diferentes titulaciones impartidas en la institución educativa.
- Programa: información sobre el programa de estudios dentro de cada titulación (acotado a modalidad a distancia).
- Malla: contiene el plan de asignaturas que viene siguiendo el estudiante.
- Periodo: indica el último periodo académico cursado por el estudiante.
- Centro: centro (núcleo) universitario de la institución educativa donde el estudiante cursa regularmente sus estudios. Al estar enfocados en modalidad a distancia, es importante analizar si la localidad del estudiante incide en la deserción
- Cantón: donde está ubicado el centro universitario
- Provincia: a la que pertenece el cantón

La figura 1 define el modelo conceptual (vista conceptual).

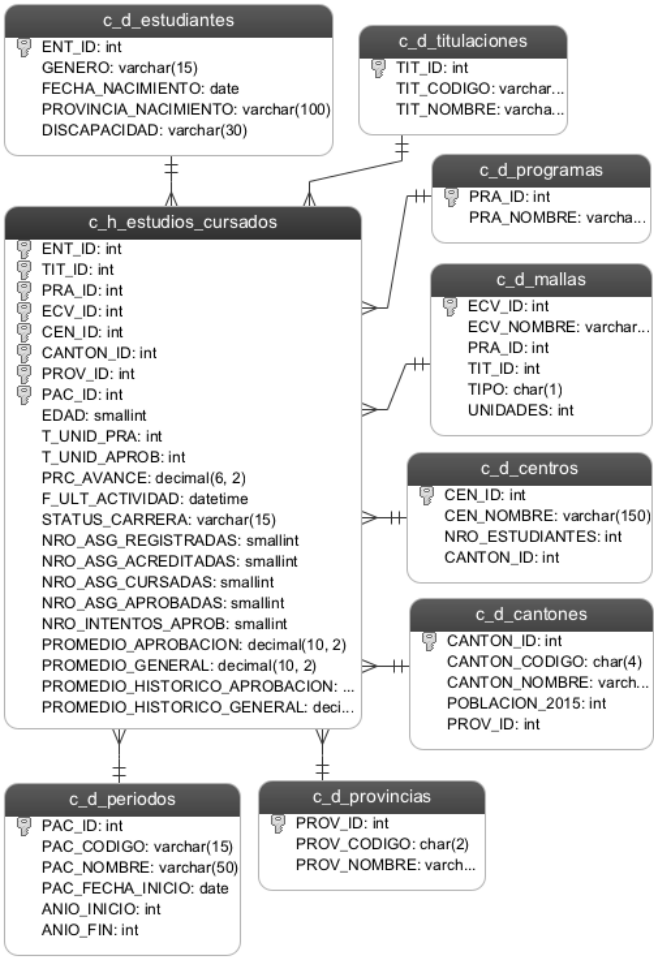

Figura 1. Modelo multidimensional (cubo "Estudios cursados")

Operaciones ETL

En esta fase definimos las operaciones específicas de (E)xtracción, (T)ransformación y (L)carga realizadas sobre las bases de datos operaciones, requeridas para construir la vista operativa a ser usada en las siguientes fases. Estas operaciones alimentarán con datos al modelo multidimensional.

Para la *extracción* se tomó como fuente la información registrada en los sistemas de gestión académica de la UTPL (institución educativa en la que se realizó la experiencia), concretamente información relativa a matrículas y expediente académico por cada estudiante que haya cursado estudios de pregrado en modalidad a distancia.

A partir de ahí se generó una vista de datos inicial de 32 atributos, proveniente de diferentes tablas de las bases de datos operacionales; dichas tablas fueron:

- Expediente: con la información sobre los centros universitarios.
- Estudiante: contiene información sobre el estudiante.
- Programa académico: información del programa académico que cursa.
- Componente educativo (asignatura): información sobre los cursos tomados.
- Resultado: contiene información sobre los resultados obtenidos.
- Periodo académico: información sobre los periodos lectivos .

En total 2,515,488 registros fueron extraídos, los cuales se analizaron, depuraron y resumieron de cara a alimentar el esquema multidimensional establecido.

Adicionalmente, información pública proporcionada por el INEC (Instituto Nacional de Estadísticas y Censos), sirvió para tener información sobre la división territorial y la población a nivel de cantones y provincias.

La *transformación* implicó resumir los datos hasta obtener un solo registro para cada estudio cursado por un estudiante, complementado con las medidas requeridas en la tabla de hechos:

- F_ULT_ACTIVIDAD: fecha estimada de última actividad en el expediente.
- EDAD: edad del estudiante la última vez que registró actividad en su expediente.
- T_UNID_PRA: número de créditos o de asignaturas requeridos para culminar los estudios.
- T_UNID_APROB: número de créditos o de asignaturas alcanzados por el estudiante.
- PRC_AVANCE: nivel de avance del estudiante.
- NRO_ASG_REGISTRADAS: total de asignaturas registradas en el expediente (incluye homologaciones).
- NRO_ASG_ACREDITADAS: total de asignaturas superadas por el estudiante (incluye homologaciones).
- NRO_ASG_CURSADAS: total de asignaturas tomadas en curso regular.
- NRO_ASG_APROBADAS: total de asignaturas aprobadas en curso regular.
- NRO_INTENTOS_APROB: total de intentos de aprobación de asignaturas tomadas en curso regular.
- PROMEDIO_APROBACION: promedio de calificaciones para asignaturas aprobadas (sobre 40 pts.)
- PROMEDIO_GENERAL: Promedio de calificaciones para asignaturas aprobadas y reprobadas.
- PROMEDIO_HISTORICO_APROBACION: promedio de aprobación histórico en la carrera para las asignaturas tomadas por el estudiante.
- PROMEDIO_HISTORICO_GENERAL: promedio de aprobación y reprobación histórico en la carrera para las asignaturas tomadas por el estudiante.
- STATUS_CARRERA: estado actual de cada estudio cursado (finalizado, en curso, abandonado).

La figura 2 muestra un ejemplo de operaciones SQL que se debieron aplicar para determinar el número de asignaturas requeridas para culminar estudios (T_UNID_PRA). En esa consulta *datos* son los datos fuente, *variaciones* es el catálogo de carreras, *pra_id* es el id de carrera, *ent_id* es el id de estudiante, *coe_id* es el id de asignatura, *tipo* es el tipo de

carrera [(A)signaturas | (C)réditos] y *etr_nombre* es el status de aprobación de la asignatura. Consultas parecidas se desarrollaron para el resto de items.

```
-- Estadistica de cantidad de estudiantes por cada
-- número de asignaturas superadas
CREATE VIEW estad_carreras_asg
FROM
SELECT c.rownum, c.pra_id, c.t_asg, c.cantidad
FROM
(
    SELECT
        IF (@fila = b.pra_id, @rownum :=@rownum + 1,
            @rownum := 1) AS rownum,
        @fila := b.pra_id AS fila,
        b.*
    FROM
    (
        SELECT a.pra_id, a.t_asg,
               COUNT(*) AS cantidad
        FROM
        (
            SELECT d.ent_id, d.pra_id,
                   COUNT(DISTINCT d.coe_id) AS t_asg
            FROM datos d, variaciones v
            WHERE d.pra_id = v.pra_id
              AND v.tipo = "A"
              AND d.etr_nombre <> "REPROBADO"
            GROUP BY d.ent_id, d.pra_id
        ) a
        GROUP BY 1,2
        ORDER BY a.pra_id, a.t_asg DESC
    )b, (SELECT @rownum:=0) x, (SELECT @fila :='') f
) c
WHERE c.rownum <= 6;

-- Estadística de número de asignaturas requeridas
-- para culminar estudios
CREATE VIEW estad_asg_requeridas
FROM
SELECT eca.pra_id,
       MIN(eca.t_asg) AS t_unid_pra
FROM
(
    SELECT ca.pra_id,
           MAX(ca.cantidad) AS cant_max
    FROM estad_carreras_asg ca
    GROUP BY ca.pra_id
) a, estad_carreras_asg eca
WHERE a.pra_id = eca.pra_id
  AND a.cant_max = eca.cantidad
GROUP BY eca.pra_id;
```

Figura 2. Operaciones SQL para indicador T_UNID_PRA

Para la *carga* de los datos al esquema multidimensional, se usó como estrategia, primero cargar los datos de los hechos, y después los asociados a sus dimensiones (sin indicadores). Finalmente, calcular los indicadores para cada hecho a medida que se calculaban.

La figura 3 muestra el proceso de cálculo y actualización del promedio de aprobación y promedio general (PROMEDIO_APROBACION, PROMEDIO_GENERAL) de cada estudio cursado. Ese proceso fue parecido para el resto de datos a cargar.

```
CREATE VIEW tmp_notas_componente_carrera
FROM
SELECT x.*, t.nota
FROM
(
SELECT DISTINCT d.ent_id, d.pra_id, d.coe_id,
       IF (d.ETR_CODIGO in
         ('HOMOLOGADO','APROBADA',
           'VALIDADO','REVALIDADO'),
           'ACREDITADO','NOACREDITADO') AS "PROMO"
FROM datos d
ORDER BY 1,3
) x, tmp_notas_componente t
WHERE x.ent_id = t.ent_id
  AND x.coe_id = t.coe_id
  AND x.PROMO = t.PROMO
ORDER BY x.ent_id, x.pra_id, x.coe_id;

CREATE VIEW estad_promedios_notas
FROM
SELECT t.ent_id, t.pra_id,
       round(ifnull(avg(IF(PROMO='ACREDITADO',
         t.nota,NULL)),0),2)
         AS "PROMEDIOAPROBACION",
       round(avg(t.nota),2) AS "PROMEDIOGENERAL"
FROM tmp_notas_componente_carrera t
GROUP BY 1,2;

UPDATE c_h_estudios_cursados h
SET  h.promedio_aprobacion =
                  (SELECT e.promedioaprobacion
                   FROM estad_promedios_notas e
                   WHERE e.ent_id = h.ent_id
                     and e.pra_id = h.pra_id)
     h.promedio_general =
                  (SELECT e.promediogeneral
                   FROM estad_promedios_notas e
                   WHERE e.ent_id = h.ent_id
                     and e.pra_id = h.pra_id);
```

Figura 3. Operaciones SQL para carga de promedios

C. Etapa 3. La extracción de conocimiento (indicadores)

En esta fase caracterizamos todo el proceso de caracterización de los indicadores y de las otras operaciones requeridas para generar conocimiento para responder a la situación objetivo (minería de datos). En ese sentido, las dos grandes tareas son: especificar las operaciones OLAP para calcular ciertos indicadores, y especificar las tareas de minería de datos.

Tipo de operaciones OLAP requeridas para el cálculo de indicadores

Tabla 1 Operaciones OLAP requeridas

Indicador	**Tipo operación OLAP**
Índice de deserción por carreras	Roll-up
Centros universitarios con mayor índice de deserción	Roll-up
Índice de deserción según rango de edad de estudiantes	Roll-up
Índice acumulado de deserción según nivel de avance	Roll-up

Por ejemplo, la figura 4 muestra la consulta roll-up específica para el indicador "Índice de deserción por carreras", y la Tabla 2 su resultado acotado a las 10 carreras con mayor índice de deserción.

```
SELECT t.tit_nombre AS "Titulacion",
       round((y.t_est_des / x.t_est),2)
       AS "IndiceDesercion"
FROM c_d_titulaciones t,
    (SELECT h.tit_id,
            count(distinct h.ent_id) AS t_est
     FROM c_h_estudios_cursados h
     GROUP BY h.tit_id) x,
    (SELECT h.tit_id,
            count(distinct h.ent_id) AS t_est_des
     FROM c_h_estudios_cursados h
     WHERE h.status_carrera = 'ABANDONADO'
     GROUP BY h.tit_id
    ) y
WHERE t.tit_id = x.tit_id
  AND t.tit_id = y.tit_id
ORDER BY 2 DESC;
```

Figura 4. Operación SQL para "Índice deserción carreras"

Tabla 2 Carreras con altos índices de deserción

Titulación	**Indice Deserción**
CIENCIAS DE LA EDUCACIÓN MENCIÓN CIENCIAS HUMANAS Y RELIGIOSAS	0.63
INFORMÁTICA	0.54
CIENCIAS DE LA EDUCACIÓN MENCIÓN QUÍMICA Y BIOLOGÍA	0.51
CIENCIAS DE LA EDUCACIÓN MENCIÓN LENGUA Y LITERATURA	0.48
INGENIERÍA EN ADMINISTRACIÓN DE EMPRESAS TURÍSTICAS Y HOTELERAS	0.48
CIENCIAS DE LA EDUCACIÓN MENCIÓN FÍSICA Y MATEMÁTICA	0.47
COMUNICACIÓN SOCIAL	0.47
ECONOMÍA	0.47
CIENCIAS DE LA EDUCACIÓN MENCIÓN INGLÉS	0.47
ADMINISTRACIÓN DE EMPRESAS	0.45

Igualmente con operaciones OLAP a partir del cubo se generó la vista de datos objeto de minería de datos (vista minable), que incluye las siguientes variables sujeto de análisis:

- Variables ya registradas en el esquema multidimensional: CodigoTitulacion, Titulacion, Genero, Edad, Centro, NroEstudiantesCentro, Canton, PoblacionCanton, Provincia, PoblacionProvincia, Tipodiscapacidad, ProvinciaNacimiento, Avance, PromedioAprobacion, PromedioGeneral, Status.
- Variables adicionales derivadas, para las cuales se diseñaron sus respectivas operaciones OLAP:
  - UltimoAño: Último año de estudios en el que se registra actividad el estudiante.
  - TasaEfectividadAprobacion: % Efectividad del estudiantes en aprobar los componentes académicos.
  - AlcanceHistoricoAprobacion: Relación porcentual del promedio de aprobación del estudiante respecto al promedio histórico de aprobación en la carrera.
  - AlcanceHistoricoGeneral: Relación porcentual del promedio de general del estudiante respecto al promedio histórico general en la carrera.

La figura 5 muestra la operación SQL aplicada sobre el cubo para generar la vista minable. Básicamente, es una operación de consulta de datos.

```
SELECT t.tit_codigo AS "CodigoTitulacion",
    t.tit_nombre AS "Titulacion",
    p.anio_fin AS "UltimoAnio",
    c.cen_nombre AS "Centro",
    c.nro_estudiantes AS "NroEstudiantesCentro",
    x.canton_nombre AS "Canton",
    x.poblacion_2015 AS "PoblacionCanton",
    p.prov_nombre AS "Provincia",
    p.poblacion_2015 AS "PoblacionProvincia",
    e.genero AS "Genero",
    UPPER(e.discapacidad) AS "Tipodiscapacidad",
    h.edad AS "Edad",
    UPPER(IFNULL(e.provincia_nacimiento,
          "<NOESPECIFICADO>"))
          AS "ProvinciaNacimiento",
    h.prc_avance AS "Avance",
    ROUND(IFNULL((h.nro_asg_aprobadas /
          h.nro_intentos_aprob),0),2)
          AS "TasaEfectividadAprobacion",
    h.promedio_aprobacion AS "PromedioAprobacion",
    h.promedio_general AS "PromedioGeneral",
    ROUND(IFNULL((h.promedio_aprobacion /
          h.promedio_historico_aprobacion),0)-1,2)
          AS "AlcanceHistoricoAprobacion",
    ROUND(IFNULL((h.promedio_general /
          h.promedio_historico_general),0)-1,2)
          AS "AlcanceHistoricoGeneral",
    h.status_carrera AS "Status"
FROM c_h_estudios_cursados h,
     c_d_estudiantes e,
     c_d_titulaciones t,
     c_d_periodos p,
     c_d_centros c,
     c_d_cantones x,
    (SELECT p1.prov_id, p1.prov_nombre,
         SUM(c1.poblacion_2015) AS "poblacion_2015"
     FROM c_d_provincias p1, c_d_cantones c1
     WHERE p1.prov_id = c1.prov_id
     GROUP BY p1.prov_id, p1.prov_nombre) p
WHERE h.ent_id = e.ent_id
  AND h.tit_id = t.tit_id
  AND h.pac_id = p.pac_id
  AND h.cen_id = c.cen_id
  AND h.canton_id = x.canton_id
  AND h.prov_id = p.prov_id;
```

Figura 5. Operación SQL Vista Minable

IV. APLICACIÓN DE LOS MODELOS DE MINERIA DE DATOS OBTENIDOS COMO PARTE DEL PROCESO DE *LA*

A. Prediccion de deserción estudiantil

Para determinar el problema de minería de datos que se debe aplicar, se analizó el contexto de la organización educativa en la cual se desarrolla la educación a distancia. Según el contexto analizado, se decidió aplicar técnicas de minería de datos para determinar el impacto de un problema en particular: "La deserción estudiantil"; la cual se da principalmente en los primeros ciclos. En este sentido, la idea no es predecir los índices de deserción, sino crear un modelo predictivo para que en base a las características conocidas de cada estudiante (atributos mencionados anteriormente), pronosticar si se trata de un alumno que potencialmente abandonará sus estudios. Dar una solución a este problema es algo de vital importancia, en particular, si mediante procesos técnicos (Data Mining) se puede determinar aquellos alumnos que potencialmente pueden abandonar sus estudios, puesto que conocer esta información con un aceptable índice de precisión puede permitir aplicar campañas de retención de estos estudiantes. Un sistema de este tipo coadyuva a la misión de la universidad, como es acompañar a estudiantes en situaciones extremas en sus procesos de enseñanza-aprendizaje, como lo es la deserción estudiantil.

Para la parte experimental fue necesario una fase de pre-procesamiento, donde se pudo determinar que el problema de deserción estudiantil se da en los niveles iniciales de estudio. En base a esto enfocamos nuestro modelo a estudiantes que estén cursando los primeros ciclos, concretamente aquellos cuyo nivel de avance sea inferior al 10%. Adicionalmente, se han eliminado ciertos atributos que no aportan a los resultados de los modelos, o tienen un alto índice de correlación con otros atributos ya incluidos, estos atributos eliminados son: "Titulacion", "UltimoAnio", "Centro", "NroEstudiantesCentro", "ProvinciaNacimiento". Asumiendo esto, procedemos con la aplicación de redes bayesianas, las cuales presentaron resultados muy aceptables; en primer lugar se trató de crear un modelo universal, es decir que pueda ser aplicado para todo nuestro universo de datos (150.000 instancias). Sin embargo, dadas las características particulares de los datos no fue posible obtener resultados aceptables con un modelo universal, por lo tanto, se decidió crear múltiples modelos aplicables a conjuntos de datos específicos, segmentados en base a determinadas condiciones. En la tabla 3 se puede observar una muestra de los resultados obtenidos. Los experimentos se han realizado utilizando un 80% de la información para entrenamiento y el 20% para pruebas.

Tabla 3 Muestra aleatoria de 8 modelos obtenidos

N.	Atributos Excluidos	Condiciones	Instancias	Precisión	Recall
1	Ninguno	2;14,12,2,1,52 | 4;5,2 | 6;1,2	2529	0.6965	0.8947
2	5	4;1,2,4 | 6;2	1637	0.7108	0.8439
3	Ninguno	4;5,2,4 | 6;1,2	4490	0.6971	0.8422
4	5	4;5,1 | 6;1	1869	0.6708	0.8278
5	**5**	**4;5,1,2 | 6;2**	**2094**	**0.8351**	**0.7403**
6	Ninguno	4;5,1,2,4 | 6;2	2555	0.8500	0.7315
7	Ninguno	2;2,52 | 4;5,2,4 | 6;1,2	1016	0.8959	0.7279
8	Ninguno	4;1,4 | 6;2	1055	0.8477	0.7193

La tabla 3 representa una muestra de 8 modelos obtenidos. Por ejemplo, el quinto modelo nos indica que se han excluido del entrenamiento 2 de los atributos explicados en la sección anterior, estos son: "Provincia" y "Población" ubicados en la posición 4 y 5, respectivamente. De la misma manera, este modelo es aplicable específicamente para los datos que cumplen dos condiciones codificadas en el algoritmo: <<4;5,1,2>> y << 6;2 >>, explicadas a continuación; cada condición tiene dos partes la primera corresponde al atributo al que se aplica el filtro y la segunda (separada por comas) corresponde a los valores que se deben mantener en ese atributo. Revisemos esto en el ejemplo anterior; la primera condición: <<4;5,1,2>> nos indica que se debe aplicar un filtro en el atributo 4 ("Provincia") y mantener los valores de 5,1,2, que corresponden a las provincias de Guayas, Loja y Azuay en Ecuador (ver Figura 6). La segunda condición << 6;2 >> nos indica que del atributo "Genero" (6) hemos seleccionado los alumnos de sexo "Masculino" (2). Así se hace para el resto de pruebas.

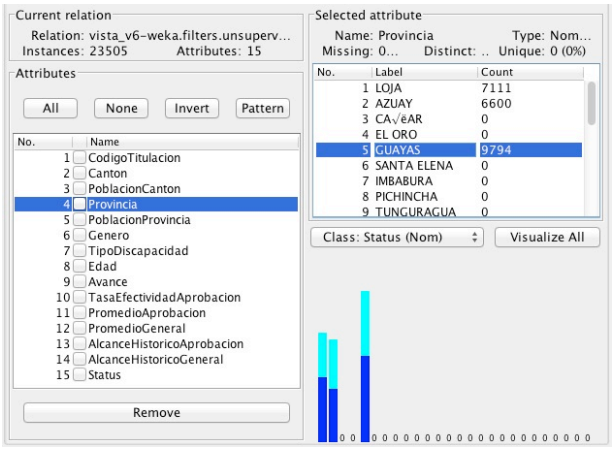

Figura 6. Entrenamiento en Weka

B. Precision y Recall

Dadas las condiciones del problema, se definieron las métricas de Precisión y Recall como las métricas adecuadas para el contexto del mismo. En específico, teniendo en cuenta que los índices de deserción estudiantil se pueden minimizar si se aplican campañas de retención y motivación a la mayoría de los estudiantes propensos a abandonar sus estudios (aunque esto implique incluir a estudiantes sin intensión de abandonarlos), consideramos que la medida de Recall debe ser la mas importante, puesto que la misma permite minimizar el índice de deserción estudiantil, ya que mide el número de ejemplos clasificados positivos correctamente entre el total de ejemplos positivos en la base de datos. Por consiguiente, se han

seleccionado aquellos modelos cuya medida de Recall es superior a 0.70 y la Precisión superior a 0.65.

El algoritmo desarrollado ha generado una gran cantidad de modelos, de los cuales 110 cumplen las condiciones de Presicion y Recall indicadas anteriormente, y cubren aproximadamente el 85% de la población estudiantil que se está estudiando.

C. Análisis de los resultados

La principal conclusión a la que podemos llegar es que no es posible conseguir un modelo universal que permita predecir la deserción estudiantil en todos los escenarios posibles, por lo que ha sido necesario establecer distintos modelos, aplicables cada uno a un contexto en particular (valores específicos de los atributos). En ese sentido, la validación de los resultados está supeditada a cada escenario.

En ese sentido, se eligieron aleatoriamente algunos de los modelos con índices de precisión y recall aceptables (superiores al 65% y 70%, respectivamente), y con ellos se aplicaron pruebas usando muestras de los datos históricos extraídos de la base de datos académica, verificando que efectivamente la predicciones de deserción son acertadas entre un 70% y 85%, con un nivel de casos no pronosticados que no supera un 30%.

La figura 7 grafica un escenario de prueba realizado con datos históricos aplicado a la carrera de ECONOMIA, cuyo índice de deserción asciende a 47% (ver Tabla 2). Se tomó una muestra de 40 estudiantes con la que se validó uno de los modelos obtenidos. Para ese escenario, 19 casos correspondieron a estudiantes desertores y 21 a no desertores. Se aplicó el modelo predictivo, y como resultado se obtuvieron 17 predicciones de deserción de las cuales 14 eran acertadas (verdaderos positivos) y 3 eran erróneas (falsos positivos). De esta manera, se obtuvo una precisión del 82.35% (14/17) y un recall del 73.68% (14/19).

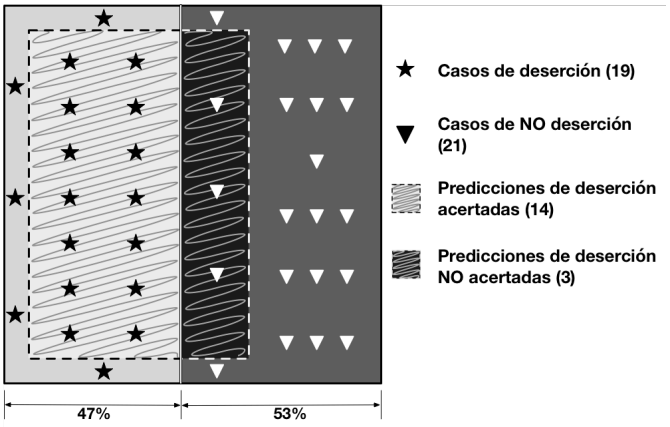

Figura 7. Escenario de prueba para carrera de ECONOMÍA

Dos aspectos fundamentales quedan por hacer:

- Se debe ampliar la vista minable operativa (añadirles más datos), para mejorar los escenarios identificados, y por consiguiente, los modelos predictivos de los mismos. Así, se contaría con una muestra más amplia de datos para la fase de entrenamiento, lo que debería redundar en mejores valores de recall y precisión para los modelos.
- Se deben empezar a usar estos modelos en un sistema gerencial de educación abierta y a distancia, con el fin de identificar potenciales casos de deserción entre sus estudiantes, sobre todo entre aquellos que cursan los primeros ciclos. Ellos permitirían a los gestores dar un seguimiento personalizado, y adoptar medidas específicas individualizadas a un estudiante o a un grupo de ellos. La efectividad y beneficio real de estos modelos predictivos, se podrá evidenciar en la medida en que los niveles de deserción de cada carrera disminuyan.

V. CONCLUSIONES

BI es una herramienta poderosa en proceso de toma de decisiones estratégicas, para reducir el tiempo y mejorar la calidad de las decisiones. BI mejora la calidad de la organización a través de un uso eficiente de los datos, con el fin de proporcionar conocimiento estratégico. Algunos de los principales riesgos en el proceso de desarrollar un proyecto de BI es una mala conceptualización de las necesidades de una organización, o el uso de datos errados. En este trabajo se trata de acotar estos problemas, a través del uso de una metodología que permite desarrollar un proceso de BI vinculado a LA, la cual contiene pasos específicos con objetivos vinculados a LA, que en nuestro caso hemos adaptado para orientarla a procesos de aprendizaje en línea centrados en el estudiante.

Las principales aportes del trabajo son: (i) definición de un proyecto de BI educativo basado en LA, (ii) diseño del proceso de desarrollo de un proyecto de BI educativo: captura de las situaciones objetivos estratégicas, diseño de los indicadores para medir su cumplimiento, etc., y (iii) verificación del proceso de desarrollo de un proyecto de BI educativo basado en LA en un caso real: procesos de aprendizaje en línea centrado en estudiantes.

De esta manera, en este trabajo se propone un enfoque para desarrollar proyectos de BI en una organización educativa, basado en LA. Ahora bien, esfuerzos futuros deberán dedicarse a analizar todas las tareas básicas posibles en un proceso de LA, aplicables a proceso de aprendizaje en línea, que permitan mejorar el desempeño estudiantil, adaptándose a sus necesidades y requerimientos (centrado en el estudiante). En este trabajo hemos obtenidos resultados preliminares muy interesantes, que motivan esos futuros trabajos, tal que permitan convertir el "conocimiento como un servicio" para optimizar procesos de enseñanza-aprendizaje.

Finalmente, la tarea de minería de datos resuelta en este trabajo responde al objetivo: Obtener el perfil del alumno potencialmente propenso a abandonar sus estudios. Aún queda por realizar una tarea extra de minería de datos para responder al objetivo: Identificar los factores que influyen en la decisión de un estudiante de abandonar sus estudios superiores.

RECONOCIMIENTO

Dr. Aguilar ha sido parcialmente financiado por el Proyecto Prometeo del Ministerio de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador.

REFERENCIAS

[1] A. Bara, I. Botha, V. Diaconita, I. Lungu, A. Velicanu, and M. Velicanu. "A model for Business Intelligence Systems', Development, Informatica Economică vol. 13, pp. 99-108, 2009

[2] M. Chatti, A. Dyckhoff, U. Schroeder and H. Thüs. "A reference model for learning analytics". International Journal of Technology Enhanced Learning (IJTEL), vol. 4, pp. 318-331, 2012

[3] M. Elbashir, P. Collier and M. Davern, "Measuring the effects of business intelligence systems: The relationship between business process and organizational performance," International Journal of Accounting Information Systems, pp. 135–153, 2008.

[4] B. Hewitt, "The online writing conference: a guide for teachers and tutors". Boynton/Cook Heinemann, Portsmouth, NJ, 2010.

[5] M. Jopling, "1:1 online tuition: a review of the literature from a pedagogical perspective," Journal of Computer Assisted Learning, vol. 28, pp 310–321, 2012

[6] O. Kozar, "The use of synchronous online tools in private English language teaching in Russia", Distance Education, vol. 33, pp. 415-420, 2012.

[7] A. Pourshahid, D. Amyot, L. Peyton, S. Ghanavati, P. Chen, M. Weiss, and A. Forster. "Business Process Management with the User Requirements Notation". Electronic Commerce Research, vol. 9, pp. 269-316, 2009

[8] P. Valdiviezo, J. Cordero. R. Reategui, J. Aguilar "A Business Intelligence Model for Online Tutoring Process", Sometido a Publicación, 2015.

[9] G. Wolfgang; D. Hendrik. "Translating Learning into Numbers: Toward a Generic Framework for Learning Analytics", Educational Technology and Society, vol. 15, pp. 42–57, 2012.