UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS
FACULTAD 4
Centro de Tecnologías para la Formación FORTES

SOLUCIÓN INFORMÁTICA PARA EL DESCUBRIMIENTO DE CONOCIMIENTO EN LOS REGISTROS DE LA PLATAFORMA EDUCATIVA NAVIGO

Trabajo presentado en opción al título de Máster en Informática Aplicada

Autor:
Ing. Osmany Montes de Oca Rodríguez

Tutores:
Dra C. Roxana Cañizares González
Dr C. Febe Ángel Ciudad Ricardo

La Habana, Cuba
2015

DEDICATORIA

A mis padres Gina y Osvaldo

A May

A todos lo que confiaron en mí

AGRADECIMIENTOS

El primer agradecimiento tiene que ser a mis padres Gina y Osvaldo, las personas que más quiero en este mundo y a quienes le debo todo lo que soy.

A Dagmay por todos su amor y apoyo en los momentos más difíciles.

A la Dra C. Roxana Cañizares González y el Dr C. Febe Ángel Ciudad Ricardo por brindarse voluntariamente a ser mis tutores, muchas gracias por su apoyo.

A Iván mi segundo padre.

A mis hermanas Olaidys y Leidy Laura por inspirarme a ser un ejemplo para ellas.

A toda mi familia por su apoyo y preocupaciones.

A mis hermanos de la vida Daimel, Yasmany y el Rafa por estar siempre pendientes.

A mi hermano de la UCI, Hector Luis por todas las batallas buenas y malas que libramos.

A Liana Isabel por brindarme su amistad y apoyo en todos estos años en la UCI.

A Lisandra Guibert por su apoyo en los momentos que lo necesité.

A toda la tropa del proyecto por la familia que logramos crear desde la generación de los tiempos de Multisaber, La caja mágica, El Navegante hasta los tiempos de Navigo y MundoClick, muy orgulloso de pertenecer a ese equipazo.

A Yunior y Gilberto por sus contribuciones con su tesis de grado.

A todos los compañeros de trabajo y conocidos que contribuyeron de alguna manera a este resultado, por sus revisiones, criterios, sugerencias o simplemente por preguntar cómo iban las cosas.

A la dirección del centro FORTES por darme el tiempo necesario para avanzar.

A la UCI por darme esta oportunidad.

A todos muchas gracias

DECLARACIÓN JURADA DE AUTORÍA

Declaro por este medio que yo Osmany Montes de Oca Rodríguez, con carné de identidad 86092511825, soy el autor principal del trabajo final de maestría "Solución informática para el descubrimiento de conocimiento en los registros de la plataforma educativa Navigo", desarrollada como parte de la Maestría en Informática Aplicada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los _____ días del mes de _____ del año _____.

_____

Ing. Osmany Montes de Oca Rodríguez

RESUMEN

En los últimos años se ha manifestado un perfeccionamiento acelerado y continuo de las Tecnologías de la Información y las Comunicaciones. La Universidad de las Ciencias Informáticas, institución que contribuye a la informatización del país, desarrolló en el centro de Tecnologías para la Formación (FORTES) la plataforma educativa Navigo, a través de la cual se gestionan hiperentornos de aprendizaje. El verdadero valor de los datos almacenados por la plataforma radica en el posible conocimiento que se puede inferir de ellos y su aplicación como apoyo a los docentes para crear informes útiles, detectar eventos, patrones, tendencias y contribuir al control y seguimiento del aprendizaje de cada estudiante o grupo en particular. Con este fin se desarrolló una solución informática que incorpora un proceso de extracción de conocimiento mediante la aplicación de minería de datos y reportes estadísticos a partir de los datos almacenados en la plataforma. El proceso implementado consta de cuatro subprocesos: Definición de los objetivos del proceso; Preparación de los datos; Minería de datos, donde se aplican las tareas de agrupamiento y reglas de asociación mediante los algoritmos K-means y Apriori y el subproceso de Análisis. Se logró la integración de la herramienta Weka con la plataforma educativa Navigo para la ejecución del proceso. Para validar la propuesta se realizaron pruebas funcionales a la plataforma y el criterio de expertos para validar los reportes obtenidos. Al concluir se pudo comprobar que la solución desarrollada contribuye al control y seguimiento de los estudiantes desde la plataforma educativa Navigo.

Palabras claves: hiperentorno de aprendizaje, minería de datos educativa, plataforma educativa Navigo

ABSTRACT

In recent years there has been a rapid and continuous development of the Information and Communications Technologies. The University of Informatics Sciences, which contributes to the computerization of the country, developed in FORTES Center the Navigo educational platform through which Learning Hyperenvironments are managed. The true value of data stored by the platform lies in the possible knowledge that can be inferred from them and their application as a support to teachers to create useful reports, identify events, patterns, trends and contribute to the control and monitoring of each student or group learning in particular. For this purpose a software solution that incorporates a process of extracting knowledge by applying data mining and statistical reports from data stored on the platform was developed. The process consists in four threads implemented: Defining the objectives of the process; Data preparation; Data mining, where the tasks of clustering and association rules by the K-mans algorithm and Apriori and thread of analysis are applied. Integrating the Navigo Weka tool platform to the execution of the process it was achieved. To validate the proposed platform, functional tests and expert judgment were conducted to validate the reports obtained. At the conclusion it was found that the developed solution contributes to the control and monitoring of students from the Navigo educational platform.

Keywords: hyperenvironment learning, educational data mining platform Navigo

ÍNDICE

INTRODUCCIÓN

En los últimos años se ha manifestado un perfeccionamiento acelerado y continuo de las Tecnologías de la Información y las Comunicaciones (TIC), en consecuencia las empresas, instituciones y gobiernos le prestan gran atención y enfocan sus estrategias en explotar los recursos que ofrecen para su desempeño, incrementándose la informatización en las diferentes esferas de la sociedad (Hilbert & Peres, 2009). El acceso al espacio virtual de la información es elemental, pero no el fin en sí mismo; es necesario poseer la correcta formación para comprender, procesar y evaluar los datos obtenidos y transformarlos en conocimiento útil (L. R. Rodríguez, 2010). En este entorno el desarrollo de software se ha convertido en un elemento de gran importancia.

Con la acentuación de las potencialidades e implantación de los sistemas informáticos también ha aumentado la cantidad de datos almacenados por los mismos. La gestión y mantenimiento de grandes cúmulos de datos supone una actividad cotidiana en muchas empresas e instituciones. La capacidad para reunir y almacenar está por encima de la capacidad para analizar y comprender conjuntos de datos masivos, por lo que ha llegado el momento en el que se dispone de tanta información que se hace difícil sacarle provecho. Los datos tal cual se almacenan (datos en bruto del inglés *raw data*) no suelen proporcionar beneficios directos; su valor real reside en la información que se puede extraer de ellos que apoye la toma de decisiones o a mejorar la comprensión de fenómenos.

El análisis de estos datos y extracción de conocimiento útil, comprensible y novedoso que está de manera oculta o implícita, no es posible mediante los métodos estadísticos convencionales (Moine, Haedo, & Gordillo, 2011). De esto derivó el nacimiento de una disciplina denominada Minería de Datos (MD), que constituye una etapa dentro del proceso de extracción de conocimiento en bases de datos, en lo adelante KDD (del inglés *Knowledge Discovery in Data Base*) (U. Fayyad, Haussler, & Stolorz, 1996; U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996b; Hernández-Orallo, Ramírez, & Ferri, 2004).

Desde los años sesenta los estadistas manejaban términos como *data fishing*, *data mining* o *data archaeology*, con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, comenzaron a consolidar los términos de MD y KDD (Félix, 2002).

Con el surgimiento del KDD los datos pasan de ser el producto generado por los diferentes procesos inherentes a la actividad desarrollada a ser la materia prima, de forma que a partir de estos volúmenes de datos se extrae conocimiento útil que ayuda a tomar decisiones. El proceso de KDD comprende diversas etapas, que van desde la obtención de los datos hasta la aplicación de conocimiento adquirido en la toma de decisiones (Hernández-Orallo et al., 2004).

La educación no está exenta del auge de las tecnologías, donde el desarrollo de software ha transformado los métodos tradicionales de enseñanza, con el objetivo de aprovechar los beneficios que brindan las TIC para un mejor desarrollo del proceso de enseñanza-aprendizaje y la obtención de resultados superiores en el ámbito educacional. Según el Dr. Pere Marquès los *"software educativos, programas educativos y programas didácticos son sinónimos para designar genéricamente los programas para ordenador creados con la finalidad específica de ser utilizados como medio didáctico, es decir, para facilitar los procesos de enseñanza y de*

aprendizaje" (Marqués, 1996; Marqués, Martinez, López, Peralta, & Zuñiga, 2009). El software educativo constituye una evidencia del impacto de las TIC en la educación, pues es una herramienta didáctica útil para estudiantes y profesores, convirtiéndose en una alternativa válida que ofrece al usuario un ambiente propicio para la construcción del conocimiento (Pérez Araujo, 2013).

Los software educativos, principalmente los basados en la web, también almacenan gran cantidad de información en sus bases de datos. Estos datos con frecuencia contienen información que puede ser de gran utilidad para profesores y estudiantes durante el proceso de enseñanza y aprendizaje. Para analizar estos datos se aplican técnicas de MD, surgiendo de esta forma la Minería de Datos Educativa, EDM (del inglés *Educational Data Mining*). Esta es una disciplina emergente considerada como una alternativa tecnológica para alcanzar otras áreas que no se tenían en cuenta en los sistemas educativos. Se basa en el desarrollo de métodos y técnicas para el análisis y la exploración de datos obtenidos particularmente desde un contexto educativo (Román, Sánchez-Guzmán, & García, 2014).

Según varios autores como (Winters, 2006), (Cristobal Romero, Ventura, Pechenizkiy, & Baker, 2010), (Peña-Ayala, 2014) y (Román et al., 2014) la EDM tiene como objetivo obtener una mejor comprensión del proceso de aprendizaje de los estudiantes y de su participación global en el mismo, orientado a la mejora de la calidad y rentabilidad del sistema educativo (Alvarez, Gonzalez, Pérez, & Espinosa, 2007).

Resulta ser realmente útil obtener datos sobre cómo los estudiantes eligen utilizar el software educativo, considerar datos a distintos niveles sobre las pulsaciones de teclas, niveles de respuestas, del alumno, de la clase o de la escuela en general. Otros temas como el tiempo, secuencia o incluso el contexto juegan papeles importantes en el estudio de datos en software educativos (Galindo & García, 2010).

En la reciente publicación del informe Horizon[1] 2015 (Johnson, Becker, Estrada, & Freeman, 2015) se hace referencia al aprendizaje y evaluación basados en datos y a las tecnologías de aprendizaje adaptativo, donde para ambas tendencias los datos y sus análisis juegan un papel importante y con esto el uso de la MD.

Algunos de los principales retos que enfrentan los investigadores en la EDM son (Morales, Soto, & Martínez, 2005):

- Integración de algoritmos de MD dentro de las propias herramientas de autor para la construcción y mantenimiento de los cursos, para la mejora automática de los sistemas.
- Desarrollo de algoritmos de MD específicos para problemas relacionados con la enseñanza y el aprendizaje utilizando entornos hipermedia adaptativos y sistemas tutores inteligentes basados en web.
- Integración de algoritmos y herramientas de MD con los entornos o plataformas educativas.

[1] Reporte donde se describen las tendencias, retos y el desarrollo de la tecnología para la educación en corto (uno o dos años), mediano (tres a cinco años) y largo plazo (cinco años o más).

- Facilidad de utilización de los algoritmos de MD. Desarrollo de herramientas fáciles e intuitivas de utilizar para los docentes que por lo general desconocen las técnicas de MD y su uso.

En el sistema de educación cubana se han introducido diferentes software educativos con el propósito de contribuir a elevar la calidad del proceso de enseñanza-aprendizaje. Ejemplos de estos son las colecciones desarrolladas por el Ministerio de Educación (MINED) como: "Multisaber" para la enseñanza primaria, "El Navegante" para las secundarias y "Futuro" para los preuniversitarios, donde todas poseen un grupo de características comunes en su diseño didáctico, resumidas en el modelo de hiperentorno de aprendizaje, definido como:

Producto basado en tecnología hipermedia que contiene una mezcla o elementos representativos de diversas tipologías de software educativo (juegos, cuestionarios, simuladores, tutoriales, glosarios, etc.), concebido para garantizar un apoyo informático a diferentes funciones del proceso docente educativo, caracterizado fundamentalmente por constituir un apoyo pleno al currículo escolar de un determinado sistema educacional (Del Toro, 2006; Gelles, del Toro, Valle, & Armenteros, 2011; Rizzo, 2009).

La Universidad de las Ciencias Informáticas[2] (UCI) es una de las instituciones que ha contribuido a la informatización del país; cuenta con varios centros de desarrollo de software donde se produce ciencia, tecnología, productos y servicios. El centro de Tecnologías para la Formación (FORTES) es uno de estos, este se especializa en la producción de aplicaciones y servicios informáticos orientados al sector educacional, para todo tipo de instituciones con diferentes modelos de formación y condiciones tecnológicas.

Entre los productos que se desarrollan en FORTES se encuentra la plataforma educativa Navigo a través de la cual se gestionan hiperentornos de aprendizaje, siguiendo la concepción didáctica de la colección de software educativo "El Navegante", diseñada por el MINED y existente en las escuelas secundarias de Cuba. La plataforma implementa el registro en Bases de Datos (BD) de todas las acciones realizadas por los usuarios durante la interacción con el hiperentorno, dígase: contenidos visitados, resultados obtenidos en los cuestionarios interactivos, tiempo dedicado a cada actividad, cantidad de visitas a cada uno de los módulos, entre otros parámetros que también se registran. Este proceso se realiza de manera cronológica por cada sesión de trabajo y de forma transparente al usuario. Se considera como una sesión cada vez que el estudiante accede al hiperentorno, por tanto si se tiene en cuenta que un estudiante puede tener varias sesiones de trabajo en el día y que una escuela puede tener un gran número de estudiantes que utilicen la plataforma, esto implica un crecimiento exponencial en el transcurso del tiempo de los datos almacenados.

Sin embargo, el verdadero valor de estos datos almacenados por la plataforma, radica en el posible conocimiento que se puede inferir de los mismos y su aplicación como apoyo a los docentes que le permita crear informes útiles, detectar eventos, patrones, tendencias, para lograr una mayor explotación de los datos y contribuir a desarrollar una mejor estrategia didáctica para la utilización del hiperentorno durante el proceso de enseñanza y aprendizaje.

Sobre lo referido en el párrafo anterior, las colecciones mencionadas desarrolladas por el MINED

[2] http://www.uci.cu/

(modelo didáctico por el que se basan los hiperentornos creados por Navigo) cuentan con un módulo denominado Resultados con el propósito de recoger información relevante para el diagnóstico del aprendizaje de los alumnos. Según estudios realizados por (L. A. R. Rodríguez et al., 2005) este módulo presenta varias deficiencias, entre las que se encuentran:

- Presentación de excesiva cantidad de información, y en ocasiones redundante, que en lugar de aportar beneficios conduce a oscurecer la lectura e interpretación de los datos.
- Se hace difícil el acceso y seguimiento de la evolución del desempeño del alumno debido a que las trazas solo se organizan por sesiones de trabajo y no por estudiantes.
- Se dificulta realizar una evaluación cualitativa o cuantitativa del desempeño ya que no ofrece de forma directa resultados o datos para esto.
- Sólo se ofrece información textual y de forma secuencial, es decir no existe un mecanismo que permita acceder directamente a las distintas secciones dentro del propio reporte.

En otras investigaciones de (L. R. Rodríguez, 2010) se plantea que: El módulo Resultados ofrece información sobre la navegación y los resultados de interacción de los alumnos con el software, muy útil para las actividades de autocontrol, autodiagnóstico, autorregulación y de comparación con respecto al grupo. Sin embargo en propia investigación plantean que la presentación de esta información aún no es lo suficientemente detallada, clara y funcional.

Otra deficiencia que se puede mencionar debido al poco aprovechamiento de los conocimientos implícitos que puedan existir en los datos almacenados, es la dificultad para que el docente pueda realizar un diagnóstico integral sobre la utilización del hiperentorno, que permita ser utilizado en la dirección del proceso de enseñanza-aprendizaje teniendo en cuenta:

- La atención a las diferencias individuales o de grupos de estudiantes con características similares en su comportamiento durante la interacción con el hiperentorno que los diferencia de otros grupos.
- La identificación de características, patrones, modelos o tendencias de los estudiantes que pueden estar implícitos en los registros.
- La detección de problemas como la falta de motivación, el bajo rendimiento, el poco uso de determinados recursos o posibles causas para los resultados evaluativos obtenidos.

De todo lo anteriormente descrito se puede concluir que:

- Existen insuficiencias en el análisis de los datos almacenados en los hiperentornos, se genera un reporte de forma textual y con gran cantidad de información que entorpece su análisis e interpretación.
- Existe dificultad para llevar a cabo un seguimiento del desempeño de los alumnos durante el uso del hiperentorno y luego realizar una evaluación cualitativa o cuantitativa del mismo.
- Se hace difícil el descubrimiento de diferencias individuales o grupos de estudiantes con características similares en su comportamiento, así como identificar características, patrones o tendencias de los estudiantes que pueden estar implícitos en los registros de los hiperentornos.

Todas estas problemáticas de forma general le dificultan a los docentes llevar a cabo un control de los resultados del aprendizaje y un seguimiento personalizado de cada estudiante o grupo en

particular.

A partir de la situación descrita hasta el momento se plantea el siguiente **problema de investigación**:

¿Cómo contribuir al seguimiento y control del aprendizaje de cada estudiante o grupo en particular a partir de los datos almacenados por la plataforma educativa Navigo?

Como **objeto de estudio** se define: la extracción del conocimiento a partir de los registros almacenados en BD.

El **objetivo general** de la investigación es: Desarrollar una solución informática que incorpore un proceso de extracción de conocimiento a partir de la aplicación de minería de datos y reportes estadísticos para contribuir al control y seguimiento personalizado del aprendizaje de cada estudiante o grupo en particular a partir de los datos almacenados por la plataforma educativa Navigo.

El **campo de acción** de la investigación está relacionado con la extracción de conocimiento a partir de técnicas de MD y reportes estadísticos en los registros almacenados por hiperentornos de aprendizaje.

Para llevar a cabo esta investigación y dar cumplimiento al objetivo propuesto, se planificaron las siguientes **tareas de investigación**:

1. Identificación de los referentes teóricos relacionando los aspectos fundamentales que sustentan la investigación mediante los cuales se consulta, extrae y recopila la información relevante sobre el descubrimiento de conocimiento en los registros almacenados por hiperentornos de aprendizaje.
2. Desarrollo de un proceso de KDD en los registros de la plataforma educativa Navigo a partir de la integración de una herramienta de MD con la plataforma.
3. Implementación de los análisis estadísticos para el módulo Resultados a partir de los datos almacenados en la plataforma educativa Navigo.
4. Validación de la solución desarrollada.

Como **hipótesis** se plantea que:

La implementación de un proceso de extracción de conocimiento a partir de la aplicación de minería de datos y reportes estadísticos contribuirá al control de los resultados del aprendizaje y seguimiento personalizado de cada estudiante o grupo en particular por parte de los docentes en la plataforma educativa Navigo.

En la investigación se destaca la utilización de los siguientes **métodos de trabajo científico:**

- Analítico-sintético: al descomponer el objeto en elementos por separado y profundizar en el estudio de cada uno de ellos, para luego sintetizarlos en la solución propuesta.
- Histórico-lógico: con el fin de realizar un estudio de los conceptos asociados a la EDM y analizar la trayectoria histórica de la MD y su incorporación en los software educativos, así como el funcionamiento y desarrollo de las técnicas más utilizadas.

- Modelación: para la representación explícita de la solución propuesta a través de la modelación del proceso de KDD, así como las ideas y referentes teóricos extraídos de las fuentes bibliográficas consultadas.
- Sistémico: para lograr que la integración entre el proceso de KDD, la herramienta de MD y la plataforma educativa Navigo funcione como un sistema.
- Criterio de expertos: para recoger la valoración por criterios de expertos sobre el resultado de la solución.

El documento está estructurado en introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas y un cuerpo de anexos. En el primer capítulo denominado **Fundamentación teórica** se analizan y exponen las teorías, enfoques, investigaciones y antecedentes que han sido considerados válidos para el desarrollo de la investigación. En el segundo capítulo que lleva por título **Propuesta de la solución** se caracteriza el proceso y luego se describen las tareas realizadas en cada una de las fases del proceso de KDD. También se describe la integración de la herramienta de MD con el hiperentorno. En el tercer y último capítulo, **Validación de la propuesta**, se expone la aplicación de los métodos de validación empleados para comprobar la validez de la hipótesis planteada.

Por último se presentan las **conclusiones** y **recomendaciones** derivadas de la investigación, una relación de las **referencias bibliográficas** y los **anexos** para facilitar la comprensión de la investigación.

CAPÍTULO 1

FUNDAMENTACIÓN TEÓRICA

"Lo que sabemos es una gota de agua; lo que ignoramos es el océano"
ISAAC NEWTON

Para establecer los fundamentos teóricos del presente trabajo es necesario analizar teorías, enfoques, investigaciones y antecedentes de gran validez para el desarrollo de la investigación. Estudiar los aspectos teóricos relacionados con la concepción pedagógica de hiperentorno de aprendizaje y los modelos más utilizados por estos para el control y seguimientos del proceso de aprendizaje de los estudiantes. Analizar las relaciones entre los conceptos MD y el proceso de descubrimiento de conocimiento en bases de datos, las diferentes fases que integran este proceso, los modelos que se pueden obtener y las diferentes tareas, algoritmos, técnicas y herramientas más utilizadas para obtenerlos. También es necesario estudiar la utilidad e integración de la MD en los entornos educativos, así como las principales herramientas para estas tareas.

1.1 Hiperentornos de aprendizaje

Este es un término muy utilizado por los pedagogos y especialistas de la informática educativa en Cuba, por lo que se ha convertido en uno de los tipos de software educativo más difundidos en desde la enseñanza infantil hasta la pre-universitaria. En (L. A. R. Rodríguez et al., 2005) se define hiperentorno interactivo de aprendizaje a las aplicaciones con fines educativos basadas en los principios de la hipermedia que integran de forma sistémica en un mismo producto a varias clases de software educativos, donde el hipertexto es el centro o un complemento. Para (Del Toro, 2006) los hiperentornos de enseñanza–aprendizaje son los sistemas hipermedia que están concebidos para ser utilizados como medios de enseñanza–aprendizaje por profesores y estudiantes. Según (Barujel, 2010) los hiperentornos constituyen un contexto, un espacio de exploración personal y grupal donde los estudiantes, a partir de su interacción con este, controlan sus actividades de aprendizaje y utilizan recursos de información y herramientas de construcción de conocimientos para resolver problemas docentes. Otro colectivo de autores (Gelles et al., 2011) definen hiperentorno como producto en el que se entremezclan diversas tipologías de software educativo (tutoriales, entrenadores, simuladores, juegos, etc.) en entornos libres hipermediales.

El concepto utilizado en esta investigación es el emitido por Labañino donde se resumen los criterios expuestos en los definiciones anteriores planteando que un hiperentorno constituye: "una *mezcla armoniosa de diferentes tipologías de software educativo sustentado en tecnología hipermedia, concebido para garantizar un apoyo informático a diferentes funciones del proceso de enseñanza aprendizaje, caracterizado fundamentalmente por constituir un apoyo pleno al currículo escolar de un determinado sistema educacional*" (Rizzo, 2009).

Este tipo de producto educativo se distingue por la integración de varias formas de presentación de la información (textos, imágenes, sonidos, animaciones, videos); posibilidades de interactividad de los estudiantes con los contenidos que le permiten el control de su aprendizaje; capacidades de almacenamiento de información sobre el trabajo realizado, entre otras características que amplían sus potencialidades didácticas (Del Toro, 2006).

Aunque no existe una estructura rígida para este tipo de productos, los que conforman las colecciones ya mencionadas como "Multisaber", "El Navegante" y "Futuro" salvo algunas diferencias originadas por la propia maduración del modelo didáctico de hiperentorno de aprendizaje o de las particularidades de los niveles de educación a los cual están dirigidos, poseen un grupo de invariantes en su diseño, las cuales se resumen en el modelo didáctico de la **figura 1**.

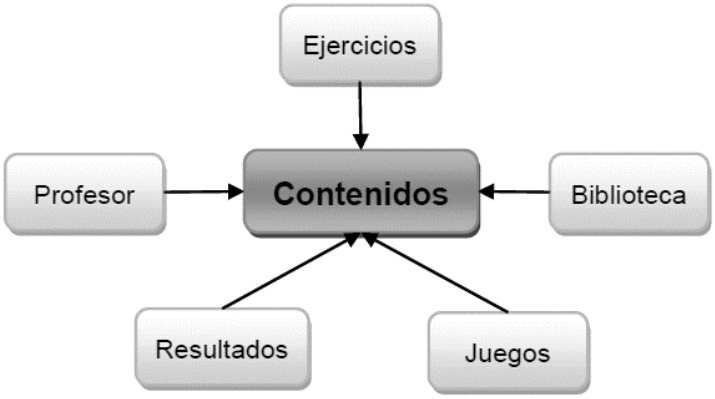


Figura 1 Modelo didáctico del hiperentorno de aprendizaje [Tomado de (L. R. Rodríguez, 2010)]

De los módulos que conforman este modelo solo es objetivo de análisis para esta investigación el módulo Resultados. Este tiene el propósito de recoger información relevante para el diagnóstico del aprendizaje de los alumnos, datos que se recogen de forma automática y transparente para el usuario durante su interacción con el hiperentorno.

1.1.1 Módulo Resultados

El modelo pedagógico de hiperentorno de aprendizaje ha evolucionado tras el paso de las diferentes colecciones ganando en madurez, supliendo deficiencias y adaptándose a las necesidades educativas de los alumnos, docentes y los niveles de enseñanza. El módulo Resultados que además de recoger toda la información de la interacción del estudiante con el hiperentorno, también propicia una vía para el control del aprendizaje, ha sido uno de los más criticado en las colecciones "Multisaber" y "El Navegante" ya que según investigaciones realizadas por (L. A. R. Rodríguez et al., 2005; L. R. Rodríguez, 2010) no es muy comprendido y por lo tanto, poco utilizado.

El módulo Resultados en los productos de estas colecciones se caracteriza por llevar un registro de cada sesión de trabajo de los estudiantes, plasmándose todos los detalles de la interacción del estudiante con el producto; es decir se registra la hora de entrada al hiperentorno, la hora de entrada a cada módulo, a cada tema, la consulta a cada recurso multimedia, los resultados de la evaluación de los cuestionarios, etc. Todo esto se realiza de manera cronológica según va sucediendo, lo que a simple vista parece ser acertado pero en la práctica genera gran cantidad de información, la cual para poseer un sentido práctico primero debe ser decodificada por el usuario, después consolidarla o resumirla para entonces obtener un criterio cualitativo o cuantitativo del desempeño del estudiante en una sesión de trabajo. Varias de las deficiencias ya fueron enunciadas en la problemática de la presente investigación.

Como se explicaba en la introducción del documento la plataforma educativa Navigo permite la gestión de hiperentornos de aprendizaje basados en la concepción de la colección "El Navegante",

pero, por todas las deficiencias que se han planteado sobre el módulo Resultados implementado en esta colección, se decide hacer un estudios sobre este módulo en otras colección para incorporar las mejores prácticas a Navigo y no arrastrar los problemas ya mencionados.

En los años 2010-2011 en la UCI se desarrolló una colección de hiperentornos llamada "La Caja Mágica" para el Ministerio del Poder Popular de la República Bolivariana de Venezuela de la cual el autor de la presente investigación fue miembro del equipo de desarrollo. Esta colección está basada en la colección cubana "Multisaber" y por tanto el módulo Resultados de esta, arrastró los problemas ya mencionados aunque sí mejoró la forma de mostrar los reportes de una forma más organizada y entendible para los estudiantes y docentes.

Por estos motivos, varios pedagogos se dieron la tarea de rediseñar el modelo de este módulo para la colección Futuro, con el fin de presentar la información sobre el desempeño del estudiante de manera que fuera relevante para la toma de decisiones por parte del profesor, y de fácil comprensión para los estudiantes como un medio más de autocontrol. Para esto decidieron continuar con los servicios de trazas e incluir otros informes que permiten un mayor análisis y sintetizado del desempeño de los estudiantes. De esta forma el módulo ofrece las siguientes opciones (C. O. C. Rodríguez et al., 2011; L. A. R. Rodríguez et al., 2005):

- **Traza del estudiante**

Con esta opción se posibilita que, una vez seleccionado el estudiante a partir de su ubicación en un grado y grupo específico además la selección de la sesión de trabajo a analizar, se pueda tener acceso a la información de todo el proceder del alumno durante su actuación con el software en la sesión escogida, a partir de un reporte que incluye, datos de la sesión (identificación, itinerario y contenidos visitados), elementos consultados (del glosario, imágenes, animaciones, diaporamas, Sonidos, Vídeos), ejercicios interactivos y juegos. El carácter de esta información es netamente descriptiva pues no se emite ningún criterio evaluativo, pero puede ser utilizada en cualquier momento por el profesor en la fase de control del aprendizaje para conocer detalles de la navegación de los estudiantes, la realización de las tareas orientadas, así como los resultados en los cuestionarios. También puede ser utilizada por el propio estudiante para autocontrolar y autorregular el aprendizaje, fomentando un aprendizaje desarrollador(L. A. R. Rodríguez et al., 2005)(L. A. R. Rodríguez et al., 2005)(L. A. R. Rodríguez et al., 2005)(L. A. R. Rodríguez et al., 2005).

- **Análisis de contenidos**

En esta opción se permite realizar un análisis de los resultados ya sea a un estudiante en particular, grupos específicos, subgrupos de estudiantes previamente seleccionados, a toda la matrícula de la escuela o a una muestra aleatoria de la misma. Este análisis se realiza sobre la base de una selección previa de los contenidos específicos sobre los que se desean realizar el estudio, obteniéndose como resultado un listado de los contenidos y el total de ejercicios respondidos en cada una de las categorías (Bien, Regular y Mal), así como una evaluación general para cada contenido. Por último se ofrece un reporte de los resultados descritos a través de gráficos estadísticos.

- **Historial del estudiante**

En esta opción se ofrece la posibilidad de que, una vez seleccionado el estudiante, a partir de su ubicación en un grado y grupo determinado, la definición de las sesiones a analizar y la selección de los contenidos sobre los que se desea realizar el estudio, se pueda tener acceso a un reporte de los resultados de este en los contenidos específicos, incluyendo un informe mediante gráficos estadísticos.

- **Análisis integral**

En esta opción, al igual que en la opción Análisis de contenidos, se permite realizar un análisis ya sea a un estudiante en particular, a grupos específicos, a subgrupos de estudiantes previamente seleccionados, a toda la matrícula de la escuela o a una muestra aleatoria de la misma, pero seleccionando las asignaturas a incluir en dicho estudio. En este análisis se obtiene como resultado un listado de las asignaturas, reportándose el total de ejercicios respondidos en cada una de ellas en las categorías de Bien, Regular y Mal, así como una evaluación general para cada una. Por último se ofrece un reporte de los resultados anteriormente descritos a través de gráficos estadísticos.

Después de este estudio se puede concluir que el diseño del módulo Resultados de la colección Futuro es el más completo con respecto a los presentes en las demás colección por lo que se decide tener en cuenta estos nuevos reportes para su implementación en la plataforma educativa Navigo. Además se concluye que este diseño tiene limitaciones para descubrir otros elementos que pueden estar implícitos en los datos, mencionadas en la problemática de la investigación, por lo que es necesario estudiar otras disciplinas como la Minería de Datos.

1.2 Definición de Minería de Datos

El término Minería de Datos surge por la aparición de nuevas necesidades y especialmente por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de instituciones, empresas, gobiernos. Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento (Hernández-Orallo et al., 2004).

Generalmente la forma más tradicional de convertir los datos en conocimiento consiste en el análisis e interpretación realizada de forma manual. Esta forma de actuar es lenta, cara y subjetiva. Este análisis manual se hace casi impracticable en dominios donde el volumen de los datos crece exponencialmente. Consecuentemente, muchas decisiones importantes se realizan siguiendo la propia intuición del usuario por no conocer o disponer de las herramientas necesarias para el análisis de los datos disponibles. "*Este es el principal cometido de la minería de datos: resolver problemas analizando los datos presentes en las bases de datos"* (Hernández-Orallo et al., 2004).

Algunos de los autores más referenciados sobre el tema definen la MD como:

- Proceso no trivial para la extracción de información implícita, previamente desconocida, y potencialmente útil desde los datos (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996a; U. Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996; Piatetsky & Frawley, 1991).
- Proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998).

- Proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos (Thuraisingham, 1998).
- Análisis de un gran conjunto de datos para encontrar relaciones desconocidas y resumir la información de forma que sean comprensibles y útiles para el usuario de los datos (Hand, Mannila, & Smyth, 2001).
- La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión (Félix, 2002).
- Exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos (Michael J Berry & Linoff, 1997; Michael JA Berry & Linoff, 2004).
- Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (I. H. Witten & E. Frank, 2005).
- Generación de conocimiento a partir de datos (Han, Kamber, & Pei, 2011).

En (Sumathi & Sivanandam, 2006) también se exponen varias definiciones de MD y al igual que en (Hernández-Orallo et al., 2004) se hace referencia a la integración de varias disciplinas donde algunas de las más influyentes son: BD, visualización, estadísticas, computación paralela, aprendizaje automático, recuperación de información y los sistemas para la toma de decisiones. En la MD además convergen diferentes paradigmas de computación como son la construcción de árboles de decisión, la inducción de reglas, las redes neuronales artificiales, el aprendizaje basado en instancias, aprendizaje bayesiano, programación lógica, algoritmos estadísticos, etc.

Después de analizar las definiciones de MD ofrecidas por varios autores se puede resumir que esta es una disciplina donde se combinan un conjunto de técnicas y algoritmos para extraer conocimiento implícito, previamente desconocido y potencialmente útil almacenado en grandes cúmulos de datos para ser utilizado con un objetivo determinado, mayormente en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos. Por tanto la tarea fundamental de la MD es descubrir modelos inteligibles a partir de los datos y el uso de estos modelos, debe ayudar a tomar decisiones más seguras que reporten beneficios para la organización.

Es significativo tener en cuenta que la utilidad del conocimiento extraído depende de la comprensión del modelo inferido, donde generalmente el usuario final no tiene por qué ser un experto en MD, como es el caso de la presente investigación donde el usuario es un maestro de la enseñanza secundaria. Por ello, es significativo hacer que el conocimiento descubierto sea mostrado de una forma comprensible (por ejemplo, usando representaciones gráficas, convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos). También es importante resaltar que la MD es solo una fase del proceso de descubrimiento de conocimiento en BD.

1.3 Minería de datos y Descubrimiento de conocimiento en BD

En el ámbito de la MD existen varios términos que se utilizan como sinónimos de esta, uno de ellos se conoce como "análisis inteligente de datos" como en (M. Berthold & Hand, 2003), donde se hace un mayor hincapié en las técnicas de análisis estadístico. Otro término muy utilizado, y el

más relacionado con la MD, es el mencionado en la introducción del documento denominado extracción o "descubrimiento de conocimiento en bases de datos". Aunque desde un punto de vista teórico o académico el término MD es una etapa dentro del KDD (U. Fayyad, Haussler, et al., 1996; U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996b; Hernández-Orallo et al., 2004), en el entorno comercial ambos términos se han utilizado indistintamente, a pesar de existir claras diferencias entre los dos. KDD pone un énfasis especial en la búsqueda de patrones comprensibles que pueden ser interpretadas como un conocimiento útil o interesante (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996b).

Por tanto el conocimiento extraído debe ser (Hernández-Orallo et al., 2004):

- Novedoso: que aporte información desconocida tanto para el sistema y preferiblemente para el usuario.
- Potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- Comprensible: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

El KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones (el objetivo de la MD), sino también la evaluación y posible interpretación de los mismos, tal y como se refleja en la **Figura 2**.

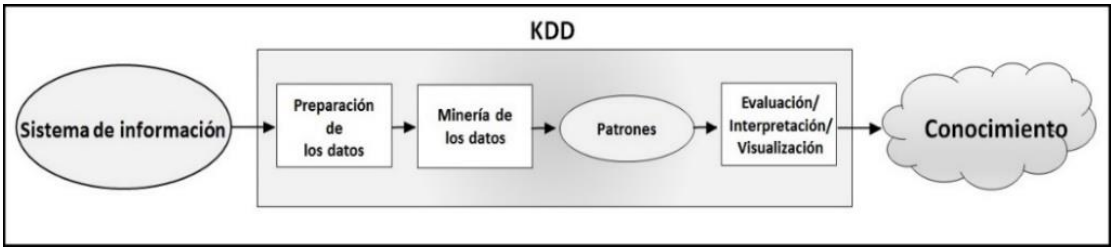


Figura 2. Proceso de KDD [Adaptado de (Hernández-Orallo et al., 2004)]

Como se puede observar el KDD permite un pre-procesamiento de los datos a partir de la selección, limpieza, transformación y proyección de los mismos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; visualizar los patrones; hacer el conocimiento disponible para su uso. A partir de esto se concluye que existe una estrecha relación entre KDD y MD: el KDD es el proceso global de descubrir conocimiento útil a partir de los datos mientras que la MD se refiere a la aplicación de los métodos para la obtención de patrones y modelos.

1.4 Clasificación de los modelos de Minería de Datos

La MD tiene como objetivo analizar los datos para extraer conocimiento. Esto puede ser en forma de relaciones, patrones o reglas inferidos a partir de los datos analizados, también pudiera ser en una descripción más concisa como un resumen. Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen varias formas de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos.

En la práctica, los modelos en función de su propósito general pueden ser de dos tipos: predictivos y descriptivos (Kantardzic, 2011). Los modelos predictivos pretenden estimar valores futuros o

desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que se denominan variables independientes o predictivas (Hernández-Orallo et al., 2004). Con frecuencia aplican funciones de aprendizaje supervisado para calcular los valores futuros o desconocidos de las variables dependientes (Hand et al., 2001). Por ejemplo, un modelo predictivo sería aquel que permite predecir el resultado final de un estudiante en un curso escolar o su permanencia en el mismo.

Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos (Hernández-Orallo et al., 2004). Suelen aplicarse funciones de aprendizaje no supervisado para producir patrones que expliquen o generalicen la estructura intrínseca, las relaciones y la interconexión de los datos extraídos (Peng, Kou, Shi, & Chen, 2008). Por ejemplo, un profesor desea identificar grupos de estudiantes con un mismo rendimiento académico, con el objeto de organizar diferentes tareas para cada grupo y poder así asignarles estas actividades; para ello analiza los resultados obtenidos anteriormente e infiere un modelo descriptivo que caracteriza estos grupos.

Según un estudio realizado por (Peña-Ayala, 2014), existe una mayor tendencia a la utilización de modelos predictivos; de una muestra de 228 trabajos en los años del 2010 al 2012, el 60% representaban modelos predictivos. Independientemente de este planteamiento, el autor de la presente investigación considera que inferir un modelo predictivo no es la solución al problema planteado.

En los hiperentornos creados con la plataforma educativa Navigo, a pesar de que tienen un carácter curricular extensivo, con el propósito de apoyar el currículo de las asignaturas, los usuarios no siguen un plan de estudios determinado donde obtengan una evaluación final, no requieren de una matrícula para usar el software. No es objetivo de la investigación predecir la evaluación final de un estudiante, su permanencia durante un curso o la rentabilidad del sistema, que es en lo que se basan la mayoría de los trabajos analizados sobre la utilización de modelos predictivos a partir de la aplicación de MD en productos educativos.

Por lo anterior descrito la presente investigación se centra en la aplicación de un modelo descriptivo que permita describir o resumir comportamientos de los estudiantes en su interacción con el hiperentorno educativo. Por tanto en lo adelante solo se abordarán tareas y técnicas de MD relacionadas con modelos descriptivos.

1.5 Fases del proceso de extracción del conocimiento

Según el diccionario de la Real Academia Española (RAE)[3], el término proceso describe la acción de avanzar o ir para adelante, al paso del tiempo y al conjunto de etapas sucesivas advertidas en un fenómeno natural o necesarias para concretar una operación artificial.

En el marco de esta investigación se utilizará la definición de proceso que propone la norma ISO 9000, donde se define proceso como: *"Conjunto de actividades mutuamente relacionadas o que interactúan, las cuales transforman elementos de entrada en resultados" (ISO9000, 2001).*

[3] http://www.rae.es/

En esta investigación se estudia el proceso de KDD. Este se organiza en diferentes fases, que como se muestran en la **Tabla 1** existe diversidad de criterios con respecto a las fases que componen el proceso de KDD pero de una forma u otra todos los autores plantean que una vez obtenido los datos se realiza una preparación de los mismos. Luego se aplican técnicas de MD para extraer los patrones y modelos, los cuales son interpretados y evaluados para la presentación del conocimiento extraído al usuario final.

Tabla 1 Fases del KDD

Autores	Fase 1	Fase 2	Fase 3	Fase 4	Fase 5
(Hernández-Orallo et al., 2004)	Integración y recopilación	Selección, limpieza y transformación	Minería de datos	Evaluación e interpretación	Difusión y uso
(Herrero & López, 2006)	Selección	Pre-procesamiento	Transformación	Minería de datos	Interpretación y consolidación
(Sumathi & Sivanandam, 2006)	Selección y limpieza	Enriquecimiento	Codificación	Minería de datos	Informes
(Sánchez, 2005)	Objetivos	Preparación de datos	Minería de datos	Análisis	Aplicación
(Morales et al., 2005)	Pre-procesamiento		Minería de datos	Post-procesamiento	
(Lara, 2011)	Selección	Pre-proceso	Transformación y reducción	Minería de Datos	Interpretación y evaluación
(Han et al., 2011)	Limpieza e integración	Selección y transformación	Minería de datos	Evaluación y presentación	

Para este trabajo se tomará como proceso de KDD el compuesto por las fases que se muestran en la **Figura 3**. Donde en la **fase de selección e integración** de datos se determina las fuentes de información que pueden ser útiles, dónde y cómo conseguirlas y en caso que proceda de distintos lugares conseguir unificarlas de manera operativa en un formato común, detectando y resolviendo las inconsistencias, dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes. Estas situaciones se tratan en la **fase de pre-procesamiento**, en la que se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos.

En la **fase de minería de datos**, se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige el método que se va a utilizar. En la **fase de evaluación e interpretación** se evalúan y analizan los patrones y si es necesario se vuelve a las fases anteriores para una nueva iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente.

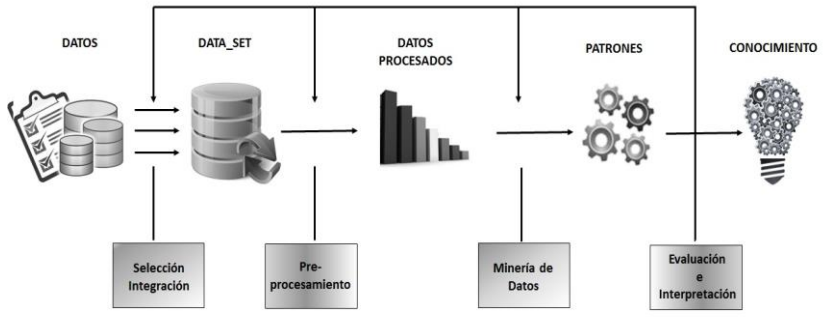


Figura 3. Fases del proceso de KD [Adaptado de (Hernández-Orallo et al., 2004)]

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema (Crows, 1999), en la que se establecen los objetivos de MD.

Existe un aspecto a tener en cuenta en el uso de KDD, es que si bien él define las fases generales del proceso de minería de datos, no especifican qué actividades puntuales hay que realizar en cada fase y deja esta definición según el criterio del equipo de trabajo.

Para la definición de las actividades dentro de cada una de las fases del proceso KDD existen varias metodologías que especifican cómo realizar cada una de ellas. Estas metodologías permiten llevar a cabo el proceso de MD en forma sistemática y no trivial; ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos (Moine et al., 2011).

Según un estudio realizado por (Moine et al., 2011) las metodologías más novedosas son SEMMA, Catalyst (conocida como P3TQ) y CRISP-DM (*CRoss Industry Standard Process for Data Mining*), siendo esta última la más utilizada en el desarrollo de proyectos de MD según la encuesta publicada en octubre del 2014 por (KDnuggets, 2014b).

CRISP-DM fue presentada en el año 1999 por las empresas SPSS, *Daimer Chrysler* y NCR (Chapman et al., 2000; Wirth & Hipp, 2000). Es una metodología abierta, desligada a ningún producto comercial y construida desde un enfoque práctico. Está estructurada en un proceso jerárquico, donde las tareas descriptas se encuentran en cuatro niveles diferentes de abstracción, van desde lo general a lo específico y tratan de abarcar la mayoría de las situaciones posibles en MD (Moine, 2013). Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación (Chapman et al., 2000). La sucesión de fases, no es necesariamente rígida. Cada una de estas fases generales se compone de un conjunto de tareas en las que se definen las salidas o entregables que se generan.

Después del estudio de la metodología CRISP-DM se pretende utilizar en esta investigación la ejecución de las actividades correspondientes a las fases que se muestran en la **Tabla 2** para la ejecución de las tareas correspondientes:

Tabla 2. Análisis de las fases y tareas de la metodología CRISP-DM [Elaboración propia]

Fase	Tareas
Comprensión del negocio	Determinar los objetivos del proceso de minería de datos
Comprensión de los datos	Recolectar los datos que se utilizarán en el proyecto y la familiarización con los mismos. • Describir los datos. • Explorar los datos. Verificar la calidad de los datos.
Preparación de los datos	Dar tratamiento de los datos para construir la vista minable o conjunto de datos final sobre el cual se aplicarán las técnicas de minería. • Seleccionar los datos • Aplicar limpieza de los datos • Construcción de los datos Integrar los datos
Modelado	Aplicar las diversas técnicas y algoritmos de minería sobre el conjunto de datos para obtener la información oculta y los patrones implícitos en ellos. • Seleccionar la técnica de modelado • Construir el modelo
Evaluación	Analizar los patrones obtenidos en función de los objetivos. Evaluar los resultados

1.5.1 Fase de selección y recopilación

En esta etapa se elige el conjunto de datos objetivo sobre los que se realizará el análisis. Una consideración importante en esta etapa es que los datos pueden proceder de diferentes fuentes y por tanto se necesitan unificarlos (Hernández-Orallo et al., 2004).

Para poder comenzar a analizar y extraer algo útil en los datos es preciso, en primer lugar, disponer de ellos. Esto en algunos casos puede parecer trivial, partiendo de un simple archivo de datos, sin embargo en otros, es una tarea muy compleja donde se debe resolver problemas de representación, de codificación e integración de diferentes fuentes para crear información homogénea (Pyle, 1999).

En estos casos lo más común es que los datos necesarios para poder llevar a cabo un proceso de KDD estén dispersos en distintas BD, tablas o incluso en diferentes locales físicamente. Esto representa un reto, ya que cada fuente de datos usa diferentes formatos de registro, diferentes grados de agregación de los datos, diferentes claves primarias, diferentes tipos de error, entre otros, por tanto lo primero es buscar una forma de integrar estos datos.

Una de las técnicas más utilizadas para esto son los llamados almacenes de datos (del inglés *datawarehousing*). Esta es uno de los métodos más aconsejables pero en algunos casos, en especial cuando el volumen no es muy grande, se puede trabajar con los datos originales o en formatos heterogéneos (archivos de texto, hojas de cálculo...) (Hernández-Orallo et al., 2004), siendo esta última opción la más aconsejable para llevar a cabo la presente investigación ya que se pueden contar con un gran cúmulo de datos pero no como para necesitar crear un almacén de datos.

1.5.2 Pre-procesamiento de los datos

La calidad del conocimiento descubierto no sólo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por ello, después de la recopilación, el siguiente paso en el proceso de KDD es seleccionar y preparar el subconjunto de datos que se va a minar, los cuales constituyen lo que se conoce como vista minable (Hernández-Orallo et al., 2004). Este paso es necesario ya que algunos datos coleccionados en la etapa anterior son irrelevantes o innecesarios para la tarea de minería que se desea realizar.

Además de la irrelevancia, existen otros problemas que afectan a la calidad de los datos. Uno de estos problemas es la presencia de valores que no se ajustan al comportamiento general de los datos (*outliers*). Estos datos anómalos pueden representar errores en los datos o pueden ser valores correctos que son simplemente diferentes a los demás.

La presencia de datos faltantes o perdidos (*missing values*) puede ser también un problema pernicioso, no obstante, es necesario reflexionar primero sobre el significado de los valores faltantes antes de tomar decisión sobre cómo tratarlos ya que éstos pueden deberse a causas muy diversas, como a un mal funcionamiento del dispositivo que hizo la lectura del valor, a cambios efectuados en los procedimientos usados durante la colección de los datos o al hecho de que los datos se recopilen desde fuentes diversas.

Otro de los pre-procesamiento importante es la selección de atributos relevantes pues es crucial que los atributos utilizados sean relevantes para la tarea de MD. También se podría construir el modelo usando todos los datos, pero si existe un número elevado se tardaría mucho tiempo y

probablemente se necesitaría un hardware más potente para su análisis. Una buena práctica para esto es usar una muestra (*sample*) a partir de algunos datos (o filas) para construir el modelo (Sánchez, 2005).

El tipo de los datos puede también modificarse para facilitar el uso de técnicas que requieren tipos de datos específicos. Así, algunos atributos se pueden numerizar, lo que reduce el espacio y permite usar técnicas numéricas. El proceso inverso consiste en discretizar los atributos continuos, es decir, transformar valores numéricos en atributos discretos o nominales. Los atributos discretizados pueden tratarse como atributos categóricos con un número más pequeño de valores. La idea básica es dividir los valores de un atributo continuo en una pequeña lista de intervalos, tal que cada intervalo es visto como un valor discreto del atributo (Hernández-Orallo et al., 2004).

Se puede concluir que es necesario poder proporcionarle a los métodos de MD el subconjunto de datos más adecuado para resolver el problema, seleccionando para estos los datos apropiados.

1.5.3 Minería de datos

La fase de MD es la más característica del KDD y es por esta razón, que en ocasiones se utiliza esta fase para nombrar todo el proceso. El objetivo de esta es producir nuevo conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas. Para ello es necesario tomar una serie de decisiones antes de empezar el proceso (Hasperué, 2012; Hernández-Orallo et al., 2004):

- Determinar qué tipo de tarea de minería es la más apropiada.
- Elegir el tipo de modelo.
- Elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que se está buscando. Esta elección es pertinente porque existen muchos métodos para construir los modelos.

Como se explica en el **epígrafe 1.3** en esta investigación se pretende obtener un modelo descriptivo y las tareas y algoritmos que se utilizarán para su construcción serán determinados en el **epígrafe 1.7**.

1.5.4 Evaluación e interpretación

En esta última etapa se evalúa y se interpreta el conocimiento extraído en la fase anterior. Aunque medir la calidad de los patrones descubiertos por un algoritmo de MD no es un problema trivial, ya que esta medida puede atañer a varios criterios, algunos de ellos subjetivos. Según las aplicaciones puede interesar mejorar algún criterio y sacrificar ligeramente otro (Hasperué, 2012; Hernández-Orallo et al., 2004).

Algunas técnicas de evaluación son (Hasperué, 2012):

- **Validación simple**: El método de evaluación más básico reserva un porcentaje de la BD como conjunto de prueba y no la usa para construir el modelo. Este porcentaje suele variar entre el 5% y el 50%. La división de los datos en estos grupos debe ser aleatoria para que la estimación sea correcta.
- **Validación cruzada con k pliegues:** En el método de validación cruzada con k pliegues (*k-fold crossvalidation*) los datos se dividen aleatoriamente en *k* grupos. Un grupo se reserva

para el conjunto de datos de prueba y con los otros *k-1* restantes se construye un modelo y se usa para predecir el resultado de los datos del grupo reservado. Este proceso se repite *k* veces, dejando cada vez un grupo diferente para la prueba. Esto significa que se calculan *k* tasas de error independientes. Finalmente se construye un modelo con todos los datos y se obtienen sus tasas de error y precisión promediando las *k* tasas de error disponibles.

- **Bootstrapping**: Otra técnica para estimar el error de un modelo cuando se disponen de pocos datos, es la conocida como *bootstrapping*. Esta consiste en construir un primer modelo con todos los datos iniciales. Se crean numerosos conjuntos de datos, llamados *bootstrap samples*, haciendo un muestreo de los datos originales con reemplazo, es decir, se van seleccionando instancias del conjunto inicial, pudiendo seleccionar la misma instancia varias veces. Nótese que los conjuntos construidos de esta forma pueden contener datos repetidos. A continuación se construye un modelo con cada conjunto y se calcula su tasa de error sobre el conjunto de test (que son los datos sobrantes de cada muestreo). El error final estimado para el modelo construido con todos los datos se calcula promediando los errores obtenidos para cada muestra.

Estas técnicas explicadas por (Hasperué, 2012) son enfocadas a modelos predictivos donde existe una clase determinada mediante la cual se puede determinar el grado de acierto del modelo, lo que se hace difícil para los modelos descriptivos donde no se cuenta con esta clase. Por lo tanto existen medidas de evaluación de modelos en dependencia de las tareas de MD utilizadas.

Agrupamiento:

Una opción que puede ser considerada es utilizar la distancia entre los grupos como medida de calidad del modelo de agrupamiento. Cuanto mayor sea la distancia entre los grupos, significa que se ha efectuado una mejor separación, y por tanto el modelo se considera mejor. Para ello es necesario determinar qué se considera como distancia entre grupos; algunas opciones pueden ser (Hasperué, 2012; Hernández-Orallo et al., 2004):

- Distancia media: la distancia entre dos grupos se calcula como la media de las distancias entre todos los componentes de ambos grupos.
- Distancia simple: se considera en este caso la distancia entre los vecinos más próximos de los dos grupos.
- Distancia completa: similar a la anterior, pero utilizando los vecinos más lejanos, es decir la distancia entre los componentes que se encuentren a más distancia.

Otra alternativa para conocer si un determinado modelo de agrupamiento es adecuado para un conjunto de datos, consiste en aprender varios modelos desde ese mismo conjunto utilizando diversas técnicas de aprendizaje. Si se comparan entre sí y coinciden, se podrá pensar que esa agrupación es acertada. Para evaluar la similitud entre dos modelos de agrupamiento se puede utilizar una estrategia, similar a la aproximación entrenamiento/ test de clasificación y regresión, basada en la partición de los datos en dos partes. La primera de ellas se utiliza para construir modelos de agrupamiento y la segunda para comprobar si los modelos construidos son similares. Para comprobar la similitud entre dos modelos *a* y *b*, con n_a y n_b grupos respectivamente, se genera una matriz con n_a filas y n_b columnas, y se inicializa a cero. Por cada dato perteneciente al conjunto de test se utilizan los modelos *a* y *b* para que asignen un grupo cada uno. Los grupos se utilizan para incrementar una celda de la matriz de comparación. Por ejemplo, si *a* retorna 1, y *b* retorna 2, se incrementa la posición (1, 2). De esta manera, tras evaluar todo el conjunto de test, si los modelos son similares, la matriz estará concentrada en unas pocas celdas. Si son diferentes, la matriz estará bastante dispersa (Hasperué, 2012; Hernández-Orallo et al., 2004).

También es posible utilizar el principio MDL[4] para evaluar modelos de agrupamiento en particular, o modelos descriptivos en general. Suponga un conjunto *E* de *n* datos, que es dividido por un algoritmo de agrupamiento en *k* grupos. Si el agrupamiento es bueno, podría utilizarse para codificar *E* de una manera más eficiente. El mejor agrupamiento permitiría la mejor manera de codificación de los datos (Hernández-Orallo et al., 2004).

Reglas de asociación:

Las reglas de asociación expresan patrones de comportamiento entre los datos. En concreto, las combinaciones de valores de los atributos (*items*) que suceden más frecuentemente. Dado que en este tipo de reglas no hay un atributo definido como clases sobre el cual evaluar la calidad de una regla, tampoco son válidos, a priori, los procedimientos vistos al inicio del epígrafe. La evaluación en las reglas se basa en los conceptos, cobertura o soporte (número de instancias a las que la regla se aplica y predice correctamente), confianza (proporción de instancias que la regla predice correctamente, es decir, la cobertura dividida por el número de instancias a las que se puede aplicar la regla) (Hasperué, 2012).

Existen otras medidas de calidad para reglas de asociación no basadas directamente en el número de ejemplos cubiertos, aunque miden otro tipo de características de las reglas tales como el interés o novedad que aporta una regla respecto al conocimiento previo (Hernández-Orallo et al., 2004).

Como conclusión, se destaca que los modelos descriptivos en general, son difíciles de evaluar debido a la ausencia de una clase determinada donde medir el grado de acierto del modelo. La mejor evaluación de este tipo de modelos es saber si el modelo resultado de la fase de aprendizaje es de utilidad en el área de aplicación.

1.6 Tareas de la Minería de Datos

Dentro de la minería existen varios tipos de tareas donde cada una puede considerarse como un tipo de problema a ser resuelto por un algoritmo de MD. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra. Las tareas pueden ser predictivas o descriptivas. Como en la presente investigación se pretende obtener un modelo descriptivo, donde los datos se presentan como un conjunto *E* sin etiquetar ni ordenar, por tanto, el objetivo no es predecir nuevos datos sino describir los existentes y para estos se pueden realizar varias tareas de MD.

Entre las tareas descriptivas se encuentran el agrupamiento (*clustering*), las reglas de asociación, las reglas de asociación secuenciales y las correlaciones. Para cumplir con el objetivo trazado en esta investigación se analizarán las tareas de agrupamiento y reglas de asociación para obtener modelos que describan el conocimiento implícito en los registros de la plataforma y de esta manera contribuir al control y seguimiento personalizado del aprendizaje de cada estudiante o grupo en particular por parte de los docentes.

[4] El principio MDL dice que la mejor descripción del conjunto de datos es aquella que minimiza la longitud de la descripción del conjunto de datos.

1.6.1 Agrupamiento

El agrupamiento (*clustering*) es la tarea descriptiva por excelencia y consiste en obtener grupos "naturales" a partir de los datos. Se habla de grupos y no de clases, porque, a diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta. El objetivo de esta tarea es obtener grupos o conjuntos entre los elementos de *E*, de tal manera que los elementos asignados al mismo grupo sean similares. Lo importante del agrupamiento respecto a la clasificación es que son precisamente los grupos y la pertenencia a los grupos lo que se quiere determinar y a priori, ni cómo son los grupos, ni cuantos hay. En algunos casos se puede proporcionar el número de grupos que se desea obtener. Otras veces este número se determina por el algoritmo de agrupamiento según la característica de los datos. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos (Hasperué, 2012; Hernández-Orallo et al., 2004). En conclusiones se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo.

La clave en este tipo de tareas es encontrar una métrica adecuada para medir las distancias entre los datos y así formar los *clusters* más acordes para una solución que resulte óptima. Si bien la distancia euclídea es una de las más utilizadas, se han propuesto otras alternativas como (Aggarwal, Hinneburg, & Keim, 2001; Ferri, Flach, & Hernández-Orallo, 2002).

Algunos de los algoritmos utilizados en las tareas de Agrupamiento son: COBWEB, EM, y K-means, DBScan, Two-step, entre otros (Garre, Cuadrado, Sicilia, Rodríguez, & Rejas, 2007).Los algoritmos K-means y DBScan basan su funcionamiento en técnicas basadas en casos, densidad o distancia. Son métodos que se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos (los casos más similares) o mediante la estimación de funciones de densidad (Hartigan & Wong, 1979). Además de los vecinos más próximos, algunos algoritmos muy conocidos son los jerárquicos, como Two-step o COBWEB. En el **epígrafe 1.7.2** se puede ver cuáles de estos algoritmos son los más utilizados en la EDM.

1.6.2 Reglas de asociación

Reglas de asociación es una tarea muy similar a las correlaciones, que tiene como objetivo identificar relaciones no explícitas entre atributos categóricos. Pueden ser de muchas formas, aunque la formulación más común es del estilo "si el atributo *X* toma el valor *d* entonces el atributo *Y* toma el valor *b*". Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados. Este tipo de tarea se utiliza frecuentemente en el análisis de la cesta de la compra, para identificar productos que son frecuentemente comprados juntos, información esta que puede usarse para ajustar los inventarios, para la organización física del almacén o en campañas publicitarias (Hernández-Orallo et al., 2004).

Un caso especial de reglas de asociación, que recibe el nombre de reglas de asociación secuenciales, se usa para determinar patrones secuenciales en los datos. Estos patrones se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre los datos se basan en el tiempo.

Algunos de los algoritmos más utilizados para el descubrimiento de reglas de asociación son: Apriori, Predictive Apriori, Apriori difuso y la familia de los algoritmos OPUS (por sus siglas en inglés: *Optimized Pruning for Unordered Search*), FP-Growth (Agraval & Srikant, 1994; Bodon,

2010; P. G. García, 2008; Hunyadi, 2011; Park, Chen, & Yu, 1995; Saldaña & Flores, 2005; Savasere, Omiecinski, & Navathe, 1995).

Entre los algoritmos descritos anteriormente, el algoritmo Apriori sobresale como el algoritmo base para la minería de reglas de asociación (Herrero & López, 2006). Este algoritmo se basa en la búsqueda de los conjuntos de *ítems* con determinada cobertura. Para ello, en primer lugar se construyen simplemente los conjuntos formados por sólo un *ítem* que supera la cobertura mínima. Este conjunto de conjuntos se utiliza para construir un conjunto de dos *ítems*, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de ítems con la cobertura requerida (Hernández-Orallo et al., 2004).

1.7 Minería de Datos Educativa (EDM)

La MD es muy utilizada en aplicaciones financieras y bancarias, análisis de mercado, comercio, seguros y salud, medicina, procesos industriales, bioingeniería, telecomunicaciones, recursos humanos, turismo, asuntos policiales, tráfico, análisis de web, entre otros. Todos estos ejemplos muestran la gran variedad de aplicaciones donde el uso de la MD puede ayudar a entender mejor el entorno donde se desenvuelve la organización y mejorar la toma de decisiones en dicho entorno.

También se ha incrementado el interés en utilizar la MD en las plataformas educativas, dando lugar a el surgimiento de la EDM como paradigma orientado a diseñar modelos, tareas, métodos y algoritmos para explotar los datos de estas plataformas, con el objetivo de desarrollar métodos de descubrimiento para hallar patrones, hacer predicciones que caractericen los comportamientos y los logros de los alumnos para permitirle al profesor una mejor comprensión de sus estudiantes y del entorno en el que aprenden (Galindo & García, 2010; Peña-Ayala, 2014).

Según los estudios realizados por (Baker & Yacef, 2009; Peña-Ayala, 2014; Peña, Domínguez, & Medel, 2009; Cristobal Romero & Ventura, 2007; Cristóbal Romero & Ventura, 2010) la EDM data de la década de los 90 pero para muchos el siglo actual representa los inicios, pues la mayoría de las referencias sobre el tema han aparecido a partir del año 2000. Actualmente existe un grupo internacional que se dedica a la temática, la Sociedad Internacional de Minería de Datos Educativa, IEDMS (del inglés *International Educational Data Mining Society)*[5]. El objetivo del grupo es dar soporte a la colaboración y el desarrollo científico en esta disciplina, a través de la organización de jornadas de trabajo, las series de conferencias EDM, la revista JEDM *(Journal of Educational Data Mining)*[6] y listas de correo, así como con el desarrollo de recursos para compartir datos y técnicas en este campo.

En consecuencia Alejandro Peña Ayala plantea que EDM está viviendo su adolescencia aunque ya las publicaciones sobre esta temática han pasado de eventos y revistas aislados a las conferencias anuales EDM[7] desde el 2008 donde también se realizan talleres y tutoriales, la edición del libro "*Handbook of educational data mining*" (Cristobal Romero et al., 2010), la revista JEDM y la ya mencionada sociedad internacional IEDMS, donde se reúnen los principales expertos en el tema.

[5] http://www.educationaldatamining.org
[6] http://www.educationaldatamining.org/JEDM/index.php/JEDM
[7] 8va Conferencia Internacional EDM http://educationaldatamining.org/EDM2015/index.php

La aplicación de técnicas de EDM se puede ver desde dos puntos de vista u orientaciones distintas (Morales et al., 2005):

- Orientado hacia los autores. Con el objetivo de ayudar a los profesores y/o autores de los sistemas de *e-learning* para que puedan mejorar el funcionamiento o rendimiento de estos sistemas a partir de la información de utilización de los alumnos. Sus principales aplicaciones son: obtener una mayor retroalimentación de la enseñanza, conocer más sobre como los estudiantes aprenden en el web, evaluar a los estudiantes por sus patrones de navegación, reestructurar los contenidos del sitio web para personalizar los cursos, clasificar a los estudiantes en grupos, etc.
- Orientado hacia los estudiantes. Con el objetivo de ayudar o realizar recomendaciones a los alumnos durante su interacción con el sistema de *e-learning* para poder mejorar su aprendizaje. Sus principales aplicaciones son: sugerir buenas experiencias de aprendizaje a los estudiantes, adaptación del curso según el progreso del aprendizaje, ayudar a los estudiantes dando sugerencias, recomendar caminos más cortos y personalizados, etc.

Después de analizar estos dos puntos de vistas se determina que para dar solución al problema planteado solo se tiene en cuenta el punto orientado hacia el profesor el cual será el encargado de llevar a cabo el proceso de KDD. Aunque con las acciones tomadas por el profesor a partir del conocimiento obtenido y la toma de decisiones el estudiante se verá beneficiado.

Hay una gran variedad de métodos empleados habitualmente en el ámbito de la educación en la MD. Estos métodos están comprendidos en las siguientes categorías: predicción, agrupamiento, minería de relaciones, inferencia a través de modelos y destilación de datos para la interpretación por parte de un ser humano. Las tres primeras categorías son universales para distintos tipos de MD (aunque en algunos casos con distintos nombres). Según (Galindo & García, 2010) las dos últimas categorías consiguen una particular importancia dentro de la EDM, principalmente el tema de hacer interpretable los modelos inferidos por personas poco conocedores en temas de MD y esto es uno de los objetivos que se pretenden lograr en esta investigación.

Una de las principales tendencias de la EDM según (Peña-Ayala, 2014) responde a la integración de un módulo EDM a la arquitectura típica de la gran diversidad de sistemas educativos basados en computadoras. También plantea que EDM ofrece varias funcionalidades durante distintas etapas del proceso de enseñanza y aprendizaje, en una primera etapa correspondiente a un apoyo proactivo con EDM para adaptar los contenidos según los perfiles de los estudiantes, en una etapa siguiente analizando los registros de la interacción estudiante-sistema para la recomendación o personalización de servicios a los usuarios en tiempo real y en una última etapa para llevar a cabo la evaluación del sistema educativo en relación a los servicios prestados, los resultados logrados, el grado de satisfacción del usuario y la utilidad de los recursos empleados.

Históricamente, ha sido difícil estudiar cómo las diferencias entre grupos de profesores o clases influencian en aspectos específicos del aprendizaje. Este tipo de análisis resulta más fácil con la MD. De manera similar, el impacto de diferencias individuales ha sido difícil de estudiar estadísticamente con métodos tradicionales. La MD aplicada al ambiente educativo posee el potencial de extender un conjunto de herramientas para el análisis de cuestiones importantes sobre diferencias individuales y grupales (Galindo & García, 2010) y precisamente esto es otro de los objetivos que se persiguen en este trabajo.

Las tareas más utilizadas en la MD aplicada a los sistemas de *e-learning* son: clasificación y agrupamiento, descubrimiento de reglas de asociación, y análisis de secuencias, aunque algunos

de los investigadores no sólo utilizan una única técnica sino la combinación de varias (Morales et al., 2005). Pero como se planteaba en el **epígrafe 1.4** en este trabajo solo se tienen en cuenta las tareas para obtener un modelo descriptivo.

Basado en el estudio realizado por (Peña-Ayala, 2014)[8], más las publicaciones analizadas sobre el tema desde el 2013 hasta el primer trimestre del año 2015, se puede mostrar a continuación algunos datos sobre las tareas y técnicas más utilizadas en EDM.

1.7.1 Tareas más utilizadas en EDM

En (Peña-Ayala, 2014) se puede observar que la tarea de MD más utilizada en EDM es la Clasificación pero que también existen una gran cantidad de trabajos que utilizan tareas de agrupamiento y reglas de asociación así como la combinación de varias. Las tareas de clasificación y agrupamiento aplicadas a sistemas de e-learning permiten agrupar a los usuarios por su comportamiento de navegación, agrupar a las páginas por su contenido, tipo o acceso y agrupar los comportamientos de navegación similares. Las reglas de asociación descubren relaciones entre atributos de un conjunto de datos que superan unos determinados umbrales. Su aplicación a sistemas de *e-learning* permite descubrir relaciones o asociaciones entre distintos recursos visitados.

En (García-Saiz, 2011) se emplean las tareas de clasificación y reglas de asociación con el objetivo de ofrecer a los profesores la posibilidad de extraer información y modelos de MD, a partir de los datos de actividades de cursos virtuales registrados en plataformas *e-learning*, sin que estos tengan conocimientos de dichas tareas y las técnicas que comprenden. En este trabajo se hace uso de los algoritmos Apriori y Predictive Apriori para la extracción de reglas de asociación. Otro ejemplo lo constituye la utilización conjunta de agrupamiento con otras técnicas como secuenciación, analizando el comportamiento de navegación de los usuarios para la personalización de sistemas *e-learning* (Mor & Minguillón, 2004).

En la bibliografía consultada se han encontrado pocos trabajos relacionados con la EDM en Cuba. Uno de ellos es el realizado en la UCI (Alvarez et al., 2007), cuyo principal resultado es el uso de técnicas de agrupamiento y asociación guiado por la metodología Crisp-DM, para determinar patrones entre la procedencia del origen social y los resultados académicos en los estudiantes de la UCI.
Varios trabajos desarrollados en el Instituto Superior Politécnico José Antonio Echeverría (CUJAE) como la tesis de maestría (Brito, 2008). Su principal objetivo es descubrir conocimiento en las bases de datos de la universidad y ayudar a la toma de decisiones de la Vicerrectoría Docente a través del uso de técnicas de MD, auxiliándose en la metodología Crisp-DM y la herramienta Weka. En dicho trabajo se emplean tareas de asociación, agrupamiento y clasificación con árboles de decisión, para agrupar a los estudiantes atendiendo a sus características, determinar cuáles son las características de los estudiantes que pueden influir en sus resultados docentes de cada año de estudio y determinar cómo influyen las características de los estudiantes en el promedio de cada año de estudio. Otros como (Sarasa, Suárez, & Sánchez, 2008; Suárez, Sarasa, & Sánchez, 2008) donde se describen las actividades realizadas siguiendo

[8] Artículo donde se analiza una muestra de 240 trabajos publicados entre el 2010 y el primer trimestre del 2013 todos relacionados con temas de EDM, donde 222 de ellos son artículos sobre los diferentes enfoques de la EDM y 18 sobre el desarrollo de herramientas para realizar EDM.

el modelo propuesto por la metodología Crisp-DM y enfocados en la herramienta de análisis Weka para la extracción de conocimientos de sus BD. Se hacen uso del algoritmo Apriori para Obtener reglas que describan las relaciones entre las características de ingreso de los estudiantes. K-means para obtener grupos de estudiantes con características de ingreso similares y árboles de decisión C4.5 para obtener, luego de la selección de atributos, cómo estos influyen en el promedio de un año de los estudiantes.

Al finalizar el epígrafe se puede concluir que las tareas seleccionadas para obtener los modelos descriptivos son utilizadas con frecuencias en la MDE. En la bibliografía consultada no se encontraron trabajos en el ámbito nacional relacionados con la aplicación de MD en hiperentornos educativos solo algunos trabajos en entornos educativos de la enseñanza superior en los cuales también se hacen usos de las tareas de agrupamiento y reglas de asociación.

1.7.2 Algoritmos más usados en EDM

El algoritmo K-means es según (Garre et al., 2007) y (Peña-Ayala, 2014) uno de los más utilizado en problemas de agrupamiento. Aunque al igual que las tareas en varios problemas la solución se basa en la combinación de varios algoritmos o la modificación de alguno de los ya existentes en dependencia de la problemática en cuestión. Con respecto a los problemas de reglas de asociación existen también una gran cantidad de algoritmos pero uno de los más usados es el Apriori, el cual constituye la base sobre la cual se han desarrollado varios algoritmos.

Para el trabajo con estos algoritmos es necesario tener conocimientos de su funcionamiento pues requieren de varios parámetros de configuración. Por ejemplo en el caso del K-means, es necesario determinar: la cantidad de *cluster* a general, que tipo de función de distancia se utilizará (distancia *Euclidean, Chevischev, Manhattan, Minkowsiki*), que estrategia seguir con los valores ausentes, cantidad de iteraciones máximas, entre otros. Con respecto al algoritmo Apriori también es necesario tener conocimientos técnicos para establecer su configuración basada en parámetros como: qué atributo se va a tomar como clase o si se van a tomar todos los atributos como clase, el soporte mínimo que las reglas deben cumplir, el tipo de métrica para determinar la calidad de la regla, el número máximo de reglas a obtener.

Después del análisis de los principales aspectos técnicos de los algoritmos más utilizados en tareas descriptivas como el agrupamiento y las reglas de asociación, se puede concluir que es necesario tener un amplio conocimiento de la materia para el trabajo con estos algoritmos. Esto resulta engorroso para usuarios con bajo o ningún dominio de la MD y sus técnicas, de ahí que sea importante ofrecer la menor cantidad de parámetros y configuraciones posibles a los usuarios.

1.7.3 Minería de Datos Educativa libre de parámetros

Uno de los principales retos que se enfrenta la EDM es el desconocimiento que los docentes tienen sobre las técnicas de MD (García-Saiz & Zorrilla, 2011). Por ejemplo, si se quiere ayudar a un profesor de una escuela secundaria que imparte la asignatura de Lengua Española o Geografía a conocer y extraer información del comportamiento de sus estudiantes con una plataforma educativa determinada, se le deben ofrecer herramientas o técnicas de MD que no requieran de configuración compleja o un entendimiento de su implementación para ser usadas. Por tanto como se plantea en la introducción, una buena práctica para dar solución a este problema es ofrecer técnicas y algoritmos de MD que no requieran parámetros iniciales o que estos los determine el algoritmo según los datos que esté tratando, lo que se conoce como MD libre de parámetros (*free parameter data mining*). Sobre este tema de la EDM libre de parámetros no se han encontrado

muchas referencias solo los artículos (Balcázar, 2011; Garcıa-Saiz & Zorrilla, 2011; Zorrilla, García-Saiz, & Balcázar, 2011) todos relacionados con la tesis de maestría (García-Saiz, 2011). A pesar de esto, es un punto importante a tener en cuenta para esta investigación, el tema de lograr que el uso de las tareas de MD sean lo más sencillo posible para los docentes.

1.8 Herramientas para la Minería de Datos

Uno de los aspectos que ha facilitado la aplicación de MD es la aparición de numerosas herramientas y paquetes, los que se pueden encontrar tanto en ámbitos comerciales como académicos, diseñados para dar soporte al ejercicio de MD. Para el análisis de estas herramientas, se tuvieron en cuenta los requisitos de la plataforma educativa Navigo y por tanto solo se analizaron las que son de código abierto, libre de pago por licencias y además las que se encuentran entre las 10 herramientas más utilizadas según la encuesta realizada por (KDnuggets, 2014a).

Knime

Knine [9] (*Konstanz Information Miner*) es un marco de trabajo gráfico para desarrollar procesos de análisis como: transformación de datos, análisis y visualización de los datos, entre otros reportes. La plataforma inicialmente desarrollada por la Universidad de Konstanz, Alemania, permite realizar muestreos, transformaciones, agrupaciones de los datos. Así como su visualización a través de histogramas. Incluye la validación de modelos a partir de algoritmos de MD tales como árboles de decisión, máquinas de soporte vectorial, regresiones, entre otros. Incluye funcionalidades adicionales como el uso de repositorios compartidos, autenticación, ejecución remota, programación, integración de SOA y una interfaz de usuario en la web. Integra todos los módulos de análisis Weka y otros plugins adicionales que permiten ejecutar scripts, que ofrece acceso a una vasta biblioteca de rutinas estadísticas y también incorpora código desarrollado en R y Python. Puede ser descargado y utilizado de forma gratuita (M. R. Berthold et al., 2008; Hasim & Haris, 2015; Knime, 2015).

RapidMiner[10]

Anteriormente conocido como YALE (por sus siglas en inglés: *Yet Another Learning Environment*) es un programa informático multiplataforma, desarrollado sobre el lenguaje Java, para el análisis y MD. Actualmente se publica bajo los términos de la Licencia AGPL. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se encuentra disponible de forma gratuita y de código abierto. Funciona en las principales plataformas y sistemas operativos y presenta un proceso muy intuitivo y con buen diseño. Puede ser accesible a través de una interfaz gráfica de usuario, modo servidor (líneas de comando) o accedido por medio de las API de Java. Posee un mecanismo de extensión simple, brinda soluciones completas. Posee un diseño de proceso gráfico para tareas estándar y un lenguaje de scripts para operaciones arbitrarias. Permite utilizar los algoritmos incluidos en Weka y el acceso a fuentes de datos, como Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase,

[9] http://www.knime.org/
[10] https://rapidminer.com/

Ingres, MySQL, Postgres, SPSS, dBase, archivos de texto y más (Hasim & Haris, 2015; Prekopcsak, Makrai, Henk, & Gaspar-Papanek, 2011; RapidMiner, 2015).

Weka

Weka (*Waikato Environment for Knowledge Analysis*)[11] desarrollado por la Universidad de Waikato, es un conjunto de librerías Java para la extracción de conocimientos desde bases de datos, que pueden ser aplicados directamente al conjunto de datos o llamados desde otras aplicaciones implementadas en Java. Está desarrollado bajo la licencia GPL (*General Public License*) e incluye interfaz gráfica compuesta por diversos entornos. Contiene una colección de herramientas de visualización, pre-procesamiento, clasificación, regresión, agrupamiento, reglas de asociación, que pueden ser accedidos a través de una interfaz gráfica. Se puede acceder a todas las funcionalidades de MD a través de una interfaz de líneas de comandos (CLI), de manera que las aplicaciones puedan sacar el máximo partido de las funciones disponibles. Permite el acceso a datos almacenados en bases de datos mediante conexión SQL. Se encuentra pública en los repositorios de los sistemas operativos Unix y por tanto se tiene acceso a su código fuentes (Hall et al., 2013; Hasim & Haris, 2015; Mikut & Reischl, 2011; Weka, 2015; I. H. Witten & E. Frank, 2005) (Report, 2010).

R

R[12] es un lenguaje y entorno de programación, multiplataforma distribuido bajo la Licencia GPL, para análisis estadístico y gráfico que brinda la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo, análisis de datos o gráficos. R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, *tests* estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas. Puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. Posee su propio formato para la documentación basado en LaTeX. Se ha desarrollado una interfaz, RWeka para interactuar con Weka que permite leer y escribir ficheros en el formato arff y enriquecer R con los algoritmos de minería de datos de dicha plataforma. R forma parte de un proyecto colaborativo y abierto. Existe un repositorio oficial de paquetes cuyo número supera los 4000, éstos se han organizado en vistas (o temas), que permiten agruparlos según su naturaleza y función (Mikut & Reischl, 2011; R-Project, 2015).

Tabla 3. Comparación de herramientas para la MD

Herramienta	Licencia	Multiplataforma	Requiere conocimientos de MD	K-means	Apriori	Lenguaje de programación	CLI
Knime	GNU GPL v3	Sí	Sí	Sí	Sí	Java	Sí
RapidMiner	GNU AGPL v3[13]	Sí	Sí	Sí	Sí	Java	Sí[14]

[11] http://www.cs.waikato.ac.nz/ml/weka/
[12] http://www.r-project.org/
[13] De código abierto (v.5 o inferior); Código cerrado (a partir de la v.6), posee una versión inicial gratis pero las *Professional*, *Personal*, *Plus* y *Enterprise* son comerciales.
[14] Solo para para la versión de servidores.

Weka	GNU GPL v2	Sí	Sí	Sí	Sí	Java	Sí
R	GNU GPL	Sí	Sí	Sí	Sí	R	Sí

Cada una de estas herramientas analizadas tiene sus ventajas y desventajas, la selección de una de ellas depende de los objetivos específicos de cada proyecto y de los conocimientos del equipo de desarrollo. En la presente investigación se selecciona a Weka como herramienta para llevar a cabo el proceso de KDD que se implementará. Según recomiendan en (Report, 2010), Weka es la herramienta ideal para aquellos usuarios que deseen desarrollar su propio software para propósitos de MD específicos y la curva de aprendizaje no es muy empinada. El equipo de desarrollo de la plataforma ya tenía conocimientos de esta herramienta, lo cual influyó en su selección. Además de las facilidades que brinda para los desarrolladores, al ser una herramienta estable, poseer una gran comunidad de desarrollo, encontrarse bien documentada y brindar soporte a las investigaciones a través de los foros y listas de correo electrónico.

Weka está compuesta por una serie de paquetes de código abierto con diferentes técnicas de pre-procesado, agrupamiento, asociación entre otras, que facilitan el trabajo, además permite el acceso a los algoritmos a través de comandos que pueden ser ejecutados desde el servidor web, lo que garantiza una fácil integración con la plataforma educativa Navigo. Otro de los aspectos que influyeron en su selección es que se encuentra publicada en los repositorios de sistemas Unix facilitando la actualización e instalación de la herramienta en el despliegue de la plataforma.

Conclusiones del capítulo

Después del análisis del estado del arte, se concluye que:

- El estudio bibliográfico asociado al objeto de estudio evidenció la importancia del uso de la MD en el descubrimiento de conocimiento, así como, el impacto significativo en los entornos educativos.
- Se evidenció la necesidad de implementar una nueva concepción del módulo Resultados de la plataforma educativa Navigo, donde se incorporen nuevos reportes y un proceso de KDD para obtener modelos descriptivos sobre el comportamiento de los estudiantes durante la interacción con la plataforma.
- Para la implementación del proceso de KDD se determinó utilizar las tareas de agrupamiento y reglas de asociación a partir de los algoritmos K-means y Apriori respectivamente, así como la integración de la herramienta de apoyo al proceso de minería Weka con la plataforma educativa Navigo.

CAPÍTULO 2
PROPUESTA DE SOLUCIÓN

Todas las verdades son fáciles de entender, una vez descubiertas. El caso es descubrirlas.
GALILEO GALILEI

Para realizar un correcto proceso de descubrimiento de conocimiento implícito en BD, es necesario pasar por una serie de fases y actividades para que los resultados finales sean relevantes. Es importante destacar que el proceso de MD es meramente exploratorio, donde se analizan los datos en búsqueda de patrones y no con el objetivo de refutar o probar la validez de un patrón determinado. En el presente capítulo se definen los reportes incorporados al módulo Resultados de la plataforma educativa Navigo así como el proceso de descubrimiento de conocimiento en los registros de la plataforma, sustentado por subprocesos, fases y un conjunto de actividades relacionadas entre ellas a través de los artefactos de entradas y salidas.

2.1 Descripción de la solución

La propuesta de solución como se muestra en la **figura 4** consiste en la incorporación al módulo Resultados de la plataforma educativa Navigo de:

- Reportes basados en la concepción didáctica diseñada por los pedagogos cubanos para los hiperentornos de aprendizajes de la colección "Futuro", como: Trayectoria del estudiante, Análisis de contenidos, Historial del estudiante y Análisis integral.
- Proceso para la extracción de conocimiento implícito en los registros almacenados en la plataforma. Este consta de varias fases y un conjunto de actividades relacionadas entre ellas a través de los artefactos de entradas y salidas. La fase fundamental de este proceso es la aplicación de las tareas de MD Agrupamiento y Reglas de asociación, mediante los algoritmos K-means y Apriori respectivamente.

Figura 4. Descripción de la propuesta [Elaboración propia]

Tanto los reportes como el proceso de KDD se implementaron con el objetivo de realizar un mayor aprovechamiento de los datos almacenados sobre la interacción de los estudiantes con los

diferentes elementos que conforman los hiperentornos educativos, para obtener un modelo descriptivo que contribuya al control y seguimiento del aprendizaje de los estudiantes desde la plataforma educativa Navigo.

2.2 Reportes descriptivos para el módulo Resultados

Cada uno de los reportes implementados para el módulo Resultados está diseñado como un asistente para su configuración, ofreciéndole al usuario la posibilidad de regresar en cualquier momento al paso anterior para cambiar una especificación. A continuación se describen en detalles cada uno de estos reportes.

2.2.1 Trayectoria del estudiante

El reporte Trayectoria del estudiante permite a los usuarios la inspección de información que se recoge y procesa producto de la interacción con el software. El carácter de esta información es netamente descriptiva pues no se emite ningún criterio evaluativo, pero puede ser utilizada en cualquier momento por el profesor en la fase de control del aprendizaje para conocer detalles de la navegación de los estudiantes por el producto, la realización de las tareas orientadas, así como los resultados en las preguntas a las que se enfrentó. También puede ser utilizada por el propio estudiante para el autocontrol y autorregulación del aprendizaje.

A partir de la selección de un estudiante se le muestra al usuario gran cantidad de información por cada una de las sesiones de trabajo registradas. El resumen se muestra por secciones como:

- **Datos generales:** Se muestra el nombre y apellido, el número de identidad, grado, grupo y escuela, así como la fecha y hora de entrada al hiperentorno.
- **Itinerario:** Muestra un resumen del recorrido realizado por el estudiante en el software, la fecha, hora de entrada y de salida, el tiempo general en el hiperentorno y detalles del itinerario como la cantidad de visitas y el tiempo en cada uno de los módulos.
- **Contenidos:** Muestra los resultados de la interacción del estudiante con el módulo a través de un resumen con la cantidad de visitas y el tiempo dedicado a cada tema, así como los detalles por cada uno de los artículos pertenecientes a los temas.
- **Ejercicios:** Muestra un resumen de la interacción con el módulo exponiendo la cantidad de ejercicios realizados, la cantidad de cada una de las categorías de revisión obtenidas, la efectividad total. Además los detalles por cada uno de los ejercicios realizados, el tiempo dedicado, nivel de complejidad, cantidad de intentos, si realizó consultas a otros módulos durante la ejercitación y el resultado obtenido.
- **Juegos:** Muestra los resultados de la interacción del estudiante con cada uno de los juegos presentes en el módulo y los detalles de los resultados según las características de cada juego en particular.
- **Galería de imágenes:** Muestra la cantidad de visitas y el tiempo dedicado a observar cada una de las imágenes visitadas.
- **Galería de videos:** Muestra la cantidad de visitas y el tiempo dedicado a observar cada uno de los videos visitados.
- **Información de interés:** Muestra los resultados de la interacción del estudiante con la información de interés que se presenta en la Mediateca del software: Curiosidades, Sabías que y otras.
- **Glosario:** Muestra los resultados de la interacción del estudiante con las palabras del glosario, tiempo y cantidad de visitas.

Con este reporte se puede conocer cuáles son los conceptos de más difícil comprensión para el estudiante en un tema determinado, donde una primera pista a seguir pudiera ser buscar en la sección Glosario los términos a los que el estudiante accedió, donde datos como la cantidad de veces que se consultó un término o el tiempo que le dedicó a su consulta pueden resultar de ayuda para los docentes.

En la **figura 5** se muestra un ejemplo del reporte Trayectoria del estudiante para la categoría de Itinerario donde se muestra toda la información sobre una sesión de trabajo seleccionada.

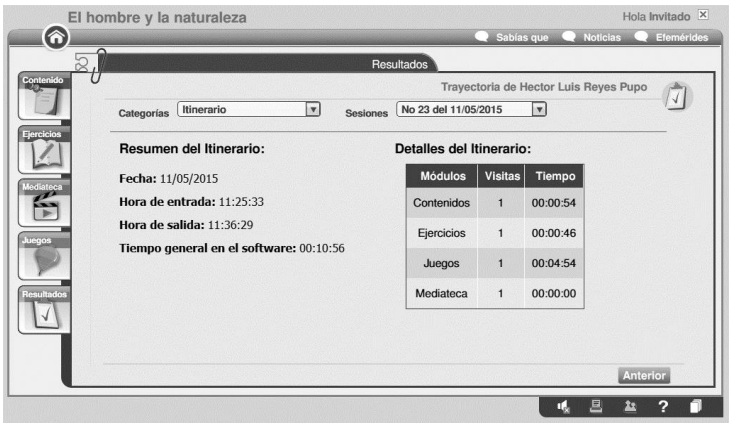


Figura 5. Trayectoria del estudiante (Categoría Itinerario)

2.2.2 Análisis de contenidos

Permite revisar la evaluación alcanzada en los cuestionarios relacionados a los contenidos específicos que se seleccionen para un estudiante, un grupo, un subgrupo, de todos los estudiantes registrados o de una muestra aleatoria del tamaño deseado (ver **figura 6**).

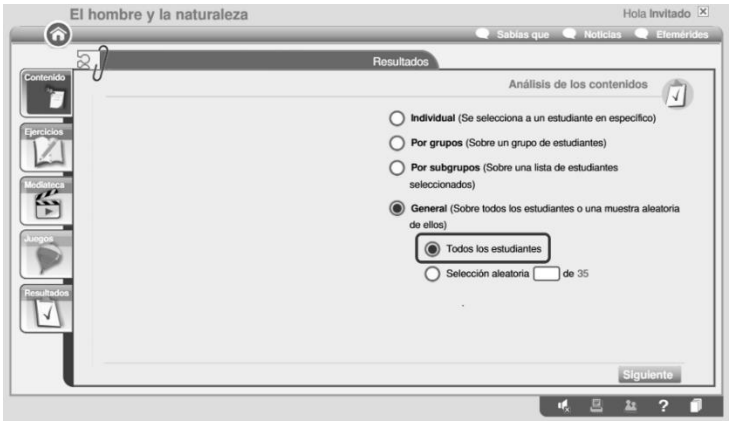


Figura 6. Selección de la cantidad de estudiantes para aplicar los reportes

Los resultados se muestran consolidados en una tabla por contenido específico en la que se recoge la cantidad de ejercicios en cada una de las categorías de revisión, la efectividad general, y la representación gráfica de esta evaluación a través de un gráfico de barras.

Este reporte (ver **figura 7**) puede ser utilizado para conocer cuál es el contenido en el que un estudiante ha presentado mayor dificultad y cuál es en el que ha obtenido mejores resultados. Pero si aumenta el ámbito del reporte se puede realizar un estudio más general a nivel de grupos escolares, de subgrupos de estudiantes que necesitan un seguimiento especial a sus diferencias

individuales, o incluso se puede utilizar esta información como parte de una investigación si se toman alumnos de forma aleatoria.

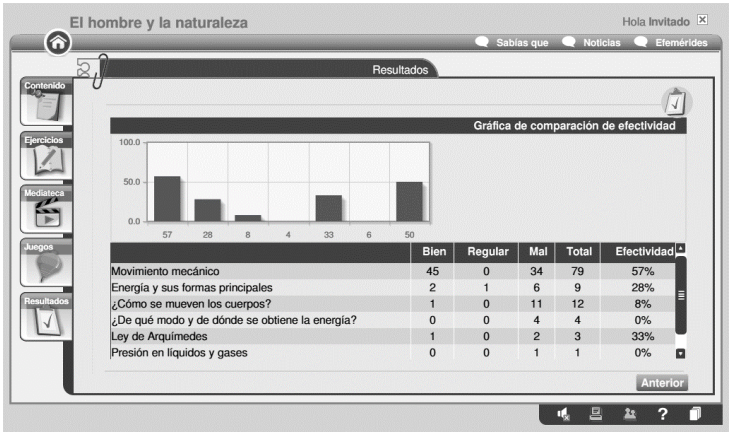


Figura 7. Análisis de contenidos

Si bien este informe fue pensado principalmente para comparar o conocer los diferentes niveles de asimilación de los contenidos por parte de los alumnos, también puede servir para comparar la eficiencia del software y del tratamiento metodológico que da al contenido, al comparar los resultados que muestra con resultados de evaluaciones realizadas por vías tradicionales.

2.2.3 Historial del estudiante

Este es un reporte (ver **figura 8**) que permite seguir a través de una línea del tiempo los resultados del estudiante en cada momento del curso o visualizarlos a través de un gráfico de líneas. Desplazándose por la línea del tiempo se puede conocer la evaluación de cada uno de los contenidos en cada sesión en la que ha trabajado el estudiante.

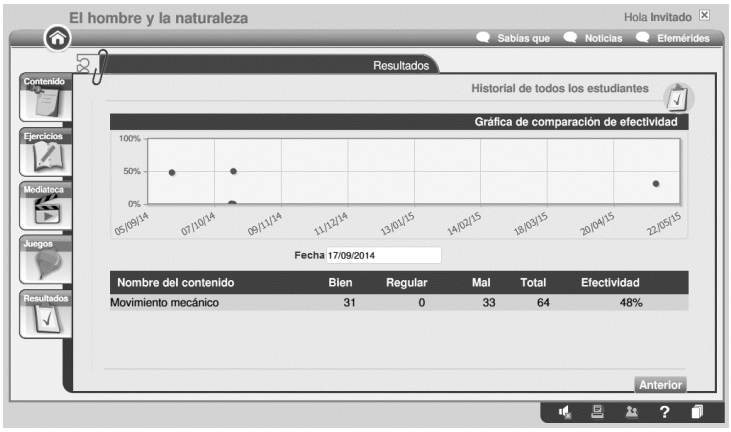


Figura 8. Historial del estudiante

2.2.4 Análisis integral

El hecho de que la plataforma permita crear una colección de hiperentornos integrados sobre un mismo núcleo y sobre una única BD, permite la oportunidad de realizar reportes que van más allá del marco de un producto o lo que es lo mismo de una asignatura. Permite dar seguimiento y control del desempeño integral del estudiante en todas las asignaturas, lo cual no es posible hacer en las colecciones existentes en las escuelas cubanas actualmente.

El Análisis integral es un reporte que permite desde cualquier producto revisar la evaluación alcanzada en cualquiera de los productos de la colección por un estudiante, un grupo, un subgrupo de estudiantes, de todos los estudiantes registrados o una muestra aleatoria del tamaño deseado. Los resultados se muestran consolidados en una tabla por un contenido específico en la que se recoge la cantidad de ejercicios en cada una de las categorías de revisión, así como los porcientos de efectividad, además se muestra un gráfico de barras con los datos de la efectividad (ver **figura 9**). Como en la BD de pruebas solo existe un hiperentorno creado llamado, "El hombre y la naturaleza" con contenidos relacionados con la asignatura de Física, es por esto que el resumen muestra una sola gráfica.

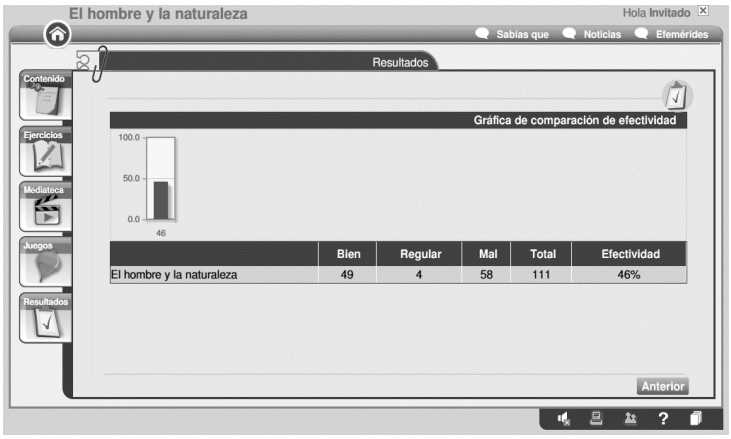


Figura 9. Análisis integral

Este tipo de reporte es de gran utilidad para que el propio alumno se percate de cuál asignatura es a la que debe dedicarle más esfuerzo y empeño, también como un medio de apoyo para el trabajo del profesor.

Todos estos reportes, Trayectoria del estudiante, Análisis de contenidos, Historial del estudiante y Análisis integral contribuyen a que los docentes y los propios estudiantes desde la plataforma educativa Navigo, lleven un control a través del tiempo de su desempeño en el aprendizaje. Además puedan identificar cuáles son las principales deficiencias de los estudiantes basadas en los resultados obtenidos a través de la interacción con el hiperentorno educativo, cuál es el contenido en el que un estudiante o un grupo de estos ha presentado mayor dificultad y cuál es el que ha sido asimilado con mayor calidad.

En las imágenes se puede observar como todos muestran de forma gráfica información que está almacenada en los registros de la plataforma y es de gran utilidad para los docentes en el control de los resultados del aprendizaje y seguimiento de cada estudiante o grupo en particular por parte de los docentes desde la plataforma educativa Navigo.

2.3 Proceso de KDD en la plataforma educativa Navigo

El proceso propuesto está compuesto por varias fases y para establecer cada una de estas se tuvo en cuenta el proceso de KDD planteado en el **epígrafe 1.5** y elementos de la metodología CRISP-DM. Este está organizado en cuatro subprocesos representados en la **figura 10**.

Inicialmente se determina el objetivo que se pretenden con la aplicación de MD a los registros almacenados en las BD de la plataforma educativa Navigo. Estos objetivos en conjunto con los registros constituyen la entrada para el subproceso Preparación de los datos.

En el subproceso Preparación de los datos se llevan a cabo varias fases como la selección de los datos que provienen de una BD relacional e integrarlos en un mismo archivo, luego realizar tareas de limpieza y transformación de los datos. Una vez que los datos estén listos se pasa al subproceso Minería de datos donde la aplican las tareas de minería, en este caso basadas en los algoritmos K-means para el agrupamiento y Apriori para la obtención de reglas de asociación. Una vez obtenido el modelo se analizan los resultados y construyen los reportes para los docentes de forma que se pueda visualizar el conocimiento. A continuación se explican en detalles cada uno de los subprocesos y las fases del proceso implementado.

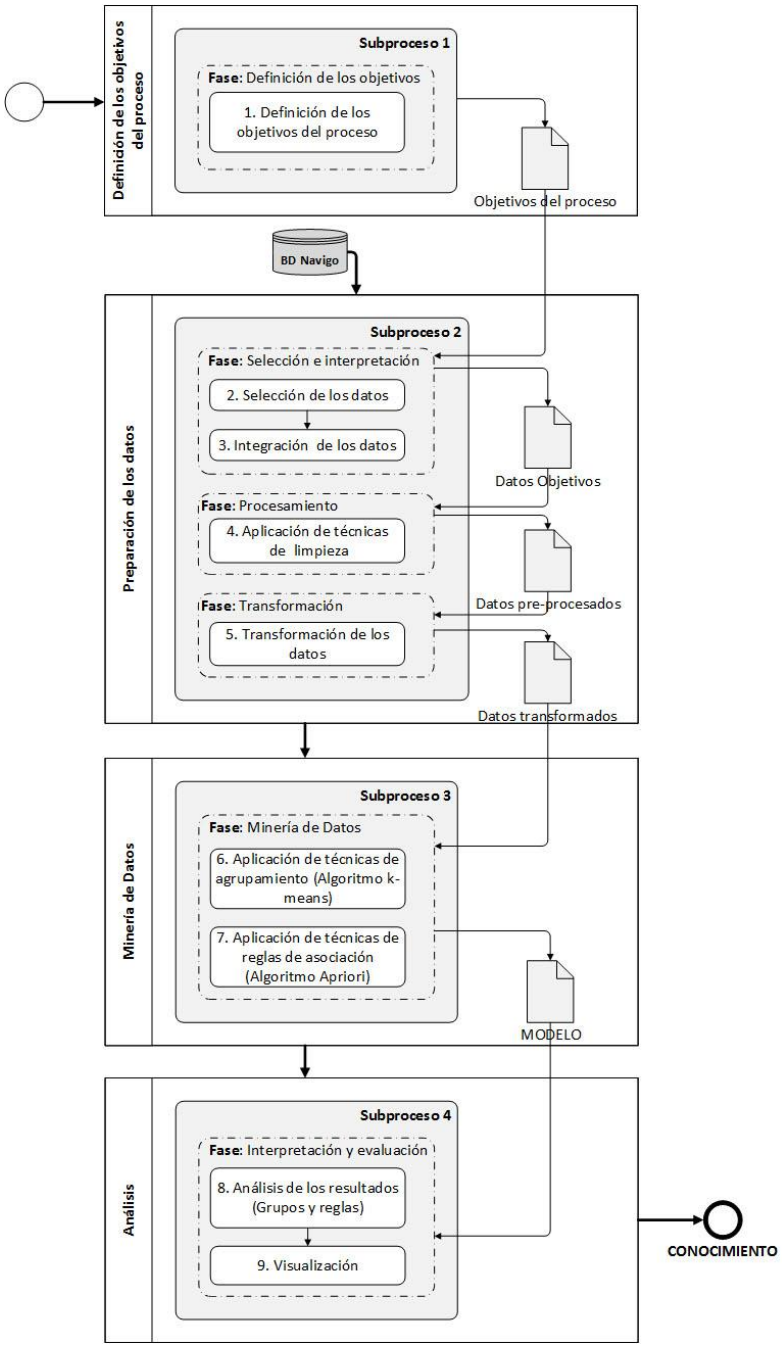


Figura 10. Descripción del proceso de KDD propuesto [Elaboración propia]

2.3.1 Objetivos

Esta etapa se centra en entender los objetivos y requerimientos del proyecto desde una perspectiva de negocio, plasmando todo esto en una definición del problema de MD basándose en las preguntas propuestas por (Microsoft, 2015):

- ¿Qué se está buscando?
- ¿Qué tipo de relaciones se buscan?
- ¿Sí se desea realizar predicciones a partir del modelo de MD o solamente buscar asociaciones y patrones interesantes?
- ¿Qué tipo de datos se tienen y qué tipo de información hay en cada columna?

Las actividades llevadas a cabo en este subproceso se muestran en la **Figura 11** y se describen a continuación.

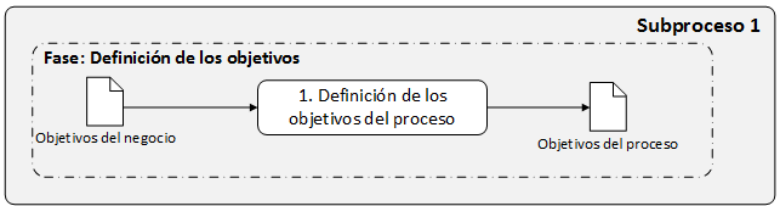


Figura 11. Descripción del Subproceso 1 [Elaboración propia]

Fase: Definición de los objetivos

Una vez entendidos los objetivos del negocio, se definen los objetivos del proceso de MD dentro del proyecto en términos técnicos y se obtiene como salida los Objetivos del proceso.

Se tiene como objetivo descubrir el conocimiento implícito en los datos almacenados en los registros de la plataforma educativa Navigo a través de la aplicación de tareas de MD como agrupamiento y reglas de asociación. Construir un reporte para el módulo Resultados que permita a los docentes identificar grupos de estudiantes con comportamientos similares durante la interacción con los hiperentornos, descubrir patrones de navegación entre otros patrones de comportamiento. Este conocimiento extraído tiene que ser útil para los docentes y expuesto de una forma clara y de fácil comprensión para cualquier tipo de usuario.

Para el desarrollo del proceso se seleccionó la herramienta Weka como base para la aplicación de los algoritmos de MD la cual está desarrollada en Java por lo que se tiene como requerimiento la instalación de la máquina virtual de Java. Ambas tecnologías son de libre uso y multiplataforma por lo que no existen restricciones en este sentido.

2.3.2 Preparación de los datos

El propósito fundamental de esta fase es manipular y transformar los datos en bruto, de manera que la información contenida en el conjunto de datos pueda ser descubierta, o más fácilmente accesible.

Subproceso 2

Este subproceso se centra en la selección y familiarización con los datos, identificar los problemas de calidad y ver las primeras potencialidades o subconjuntos de datos que puedan ser interesantes para analizar luego del procesamiento y trasformación de estos como parte de la extracción del

conocimiento. Las actividades llevadas a cabo en este subproceso se muestran en la **Figura 12** y se describen a continuación.

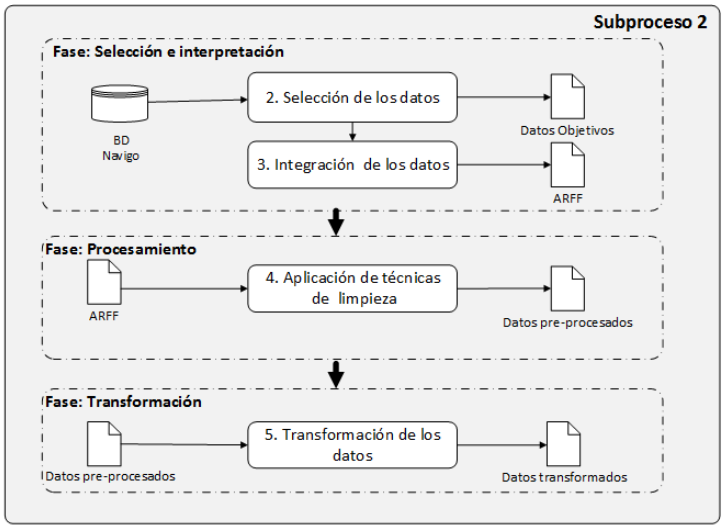


Figura 12 Descripción del Subproceso 2 [Elaboración propia]

Fase: Selección e integración

Una vez determinado los objetivos a los que se quieren llegar con la aplicación de la MD, es necesario seleccionar los datos. Por esta razón, serán la entrada base al proceso de descubrimiento y por tanto a esta primera etapa de Selección. Esta etapa tiene como objetivo la preparación de las fuentes de datos y la selección de las mismas.

Actividad 2: Selección de los Datos

Los hiperentornos cuentan con seis módulos (Contenido, Ejercicios, Mediateca, Juegos Resultados y General) que pueden ser accedidos por los estudiantes. Resulta interesante analizar la interacción de estos con cada uno de los módulos (exceptuando Resultados y General), con el fin de encontrar patrones de navegación y grupos de comportamientos similares entre los estudiantes en el uso del hiperentorno. También pueden descubrirse otros aspectos importantes a tener en cuenta por los profesores a la hora de estructurar los elementos que integran la colección y guiar el uso del producto educativo. Para esto se realizó un estudio de las diferentes tablas en la DB para analizar qué información se almacena por cada uno de los módulos. Después de esto se cuenta con los siguientes datos:

- Contenido: Cantidad de visitas a cada tema y a sus artículos, así como el tiempo dedicado en cada visita.
- Ejercicios: Por cada ejercicio realizado, el nivel de complejidad, el tiempo dedicado, la cantidad de intentos consumidos, si lo realizó en equipo, si realizó consultas, la tipología y la evaluación. Además se cuenta con la efectividad por cada uno de los temas.
- Juegos: Cantidad de visitas al módulo, la efectividad de las soluciones en los juegos como el Crucigrama, la Sopa de letras y el Descubre la imagen además de los niveles de complejidad.
- Mediateca: Cantidad de visitas a cada uno de los elementos que conforma la Mediateca como son: galerías de imágenes, de videos, glosario de términos, artículos de interés, personalidades.

- Datos de cada uno de los estudiantes, como el nombre y sus apellidos, el grupo, grado y escuela a la que pertenece.

De todo este cúmulo de datos, teniendo en cuenta el aporte que pueden brindar para el control y seguimiento del desempeño de los estudiantes, se decide seleccionar para el proceso de KDD los datos relacionados con:

- Visitas a cada uno de los módulos y el tiempo de la misma.
- Evaluaciones obtenidas en los ejercicios y la efectividad general.

Actividad 3: Integración de los Datos

Los datos seleccionados se encuentran en una BD relacional como se puede observar en el **ANEXO 2**. Esto implica que gran parte de la información se encuentre en las relaciones entre las tablas, lo cual dificulta el procesamiento de los datos a través de la herramienta Weka. Por lo anterior se hace necesario extraer los datos y almacenarlos en una estructura con la que Weka pueda interactuar. Esta establece varias formas para definir el origen de los datos, que puede ser estableciendo la ruta y los parámetros de conexión a una BD no relacional, a través de ficheros binarios, CSV (del inglés *comma-separated values*) [15] o ARFF (del inglés *Attribute-Relation File Format*). Para esta investigación el método seleccionado es el nativo de Weka, el archivo ARFF.

El fichero ARFF está compuesto fundamentalmente por tres partes: cabecera, declaración de atributos y sección de los datos.

En la cabecera se define el nombre de la relación siguiendo como estructura:

```
@relation <nombre_relación>
```

En la sección de declaración de los atributos se declaran todos los atributos que contendrá el archivo, unido al tipo de datos del atributo de la siguiente forma:

```
@attribute <nombre_atributo> <Tipo_atributo> o <rango>
```

Los tipos de atributos que define Weka son:

String: Para expresar cadenas de texto.

Nominal: Para expresar los posibles valores que puede alcanzar una variable. Se enuncian entre llaves y separados por coma.

Numeric: Para expresar los números reales.

Integer: Para expresar los datos que representan números enteros.

Date: Expresa fechas, para ello debe ir precedido de una etiqueta de formato entrecomillada la cual está compuesta por caracteres separadores y unidades de tiempo: dd,MM,yyyy,HH,mm,ss.

[15] Son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal: Argentina, Brasil...) y las filas por saltos de línea.

La última sección es donde se declaran las instancias que forman la relación. El inicio de la sección viene dado de la forma `@data`. Los atributos que forman las instancias se separan por coma y las instancias por saltos de líneas.

En la **figura 13** se muestra un fragmento del fichero ARFF con alguno de los datos seleccionados cumpliendo con la estructura propuesta por Weka.

```
@RELATION  Themes1ExerciseEffectiveness

@ATTRIBUTE  studentID Integer
@ATTRIBUTE  numVisitTheme1 Integer
@ATTRIBUTE  timeTheme1 Date "HH:mm:ss"
@ATTRIBUTE  numExeTheme1 Integer
@ATTRIBUTE  numExeTheme1B Integer
@ATTRIBUTE  numExeTheme1R Integer
@ATTRIBUTE  numExeTheme1M Integer
@ATTRIBUTE  effectivExeTheme1 real

@DATA
103,10,01:55:56,40,30,5,5,95.3
104,40,02:30:43,48,35,10,3,90.3
105,23,01:55:56,40,30,5,5,95.3
106,5,00:10:56,37,16,15,6,70.8
110,10,01:15:51,40,31,7,2,95.7
99,10,01:55:56,40,38,2,0,99.55
101,10,01:10:21,40,30,5,5,95.3
102,3,00:05:58,10,2,5,3,30.3
107,10,01:55:56,40,30,5,5,95.3
108,10,01:55:56,40,30,5,5,95.3
109,10,01:55:56,40,30,5,5,95.3
111,10,01:55:56,40,30,5,5,95.3
```

Figura 13. Ejemplo de archivo .arff

Es importante aclarar que como la plataforma no está en explotación aún los datos con los que se experimenta en el proceso son datos de prueba para poder implementar y comprobar las funcionalidades del proceso.

Los datos objetivos integrados en el archivo ARFF constituyen las salidas de esta actividad y las entradas para la fase de Pre-procesamiento.

Fase: Pre-procesamiento

La fase Pre-procesamiento contribuye a garantizar la calidad de la información sobre la que se pretende extraer conocimiento antes de llegar a aplicar las técnicas de MD. Esto se debe a que entre mayor sea esta calidad, mayor será la calidad de los modelos generados a partir de dicha información y que pueden utilizarse en la toma de decisiones. En este sentido, la obtención de información útil para ser posteriormente procesada es un factor clave (E. P. Rodríguez, 2014).

En la creación del ARFF se debe revisar el nivel de calidad de los datos que la conforman. El trabajo inadecuado con información y la unión de datos de distintas fuentes pueden generar anomalías en los datos. Estas anomalías suelen provocar inferencias erróneas sobre los datos almacenados y por tanto que se tomen decisiones incorrectas. Para evitar este tipo de problemas se realizan técnicas de limpieza de datos como parte de la etapa de Pre-procesamiento.

Actividad 4: Aplicación de técnicas de limpieza

El objetivo de esta actividad es facilitar y simplificar el problema en cuestión, sin excluir o dañar información importante para el proceso de modelado. La reducción de espacio de entrada es un proceso de suma importancia desde el punto de vista del rendimiento de la aplicación.

Una de las operaciones que se realizan en la fase de limpieza de los datos son el trabajo con los valores omisos y la eliminación de los datos incorrectos. Con respecto a esto último, los datos seleccionados para ser analizados son obtenidos a partir de la interacción de los usuarios con el sistema donde estos no introducen datos, ya que todos los datos son basados en los clic que realizan los usuarios dentro del hiperentorno. Esto posibilita que no se introduzcan datos incorrectos. Relacionado con los valores omisos, por lo planteado anteriormente en los registros de la plataforma no existen valores en blanco y por lo tanto no se tienen en cuenta.

Como salida de esta actividad se obtienen los datos pre-procesados.

Fase: Transformación

La fase de Transformación en la extracción del conocimiento se encuentra estrechamente relacionada con la fase de Pre-procesamiento. Esto se debe a que para transformar los datos y aplicar técnicas en el subproceso Minería de Datos, deben haberse pre-procesado los datos anteriormente. La actividad que se realiza en esta etapa es la transformación de los datos.

Actividad 4: Transformación de los datos

Es cierto que existen varios tipos de datos (enteros, reales, fechas, cadenas de texto, etc.) y como se explicó en la **Actividad 3**, Weka brinda soporte para diferentes tipos de datos, pero desde el punto de vista de las técnicas de MD más habituales solo interesa distinguir entre dos tipos: numéricos (enteros o reales) y categóricos o discretos (toman valores en un conjunto finito de categorías). Incluso considerando sólo estos dos tipos de datos, se debe aclarar que no todas las técnicas son capaces de trabajar con ambos tipos.

Ejemplo de lo anterior es el algoritmo seleccionado para la tarea de agrupamiento, el K-means solo es aplicable sobre datos numéricos, por tanto es necesario transformar los atributos no numéricos que sean significativos para el proceso de modelado en valores numéricos. Ejemplo de esto son los atributos de tipo *Date*, los cuales tiene un formato "HH:mm:ss" se trasformó el valor en *Integer* después de convertir el valor del tiempo a segundos.

Ejemplo:

```
HH:mm:ss = valor Integer
01:55:56 = 6956
```

Para el algoritmo Apriori es necesario convertir todos los atributos numéricos en nominales, para lo cual se aplicó el filtro *Discretize*.

```
weka.filters.unsupervised.attribute.Discretize
```

Este consiste en transformar los atributos numéricos seleccionados en atributos simbólicos, con una serie de etiquetas que resultan de dividir la amplitud total del atributo en intervalos (Amaya Torrado, Barrientos Avendaño, & Heredia Vizcaíno, 2014; Jiménez & Álvarez, 2010). Si bien la discretización de una variable puede ser realizada de manera arbitraria visualizando los valores de esa variable, se han presentado trabajos que realizan esta tarea de manera automática (Hu, Chen, & Tang, 2009; Hua & Zhao, 2009). Una vez ejecutado el filtro se obtiene un resultado como el siguiente:

```
@attribute studentID numeric
@attribute visitThemes {'\'(-inf-68.333333]\'','\'(68.333333-120.666667]\'','\'(120.666667-inf)\''}
@attribute visitImagery {'\'(-inf-20]\'','\'(20-35]\'','\'(35-inf)\''}
@attribute visitVideos {'\'(-inf-20]\'','\'(20-35]\'','\'(35-inf)\''}
@attribute visitGlossary {'\'(-inf-16.666667]\'','\'(16.666667-28.333333]\'','\'(28.333333-inf)\''}
@attribute visitGames {'\'(-inf-20]\'','\'(20-35]\'','\'(35-inf)\''}
@attribute numExeTheme {'\'(-inf-15.333333]\'','\'(15.333333-27.666667]\'','\'(27.666667-inf)\''}
@attribute effectivExeTheme1 {'\'(-inf-60]\'','\'(60-80]\'','\'(80-inf)\''}
@attribute numVisitTheme2 {'\'(-inf-14]\'','\'(14-27]\'','\'(27-inf)\''}
@attribute effectivExeTheme2 {'\'(-inf-62.5]\'','\'(62.5-81.25]\'','\'(81.25-inf)\''}
@attribute numVisitTheme3 {'\'(-inf-15.333333]\'','\'(15.333333-27.666667]\'','\'(27.666667-inf)\''}
@attribute effectivExeTheme3 {'\'(-inf-62.5]\'','\'(62.5-81.25]\'','\'(81.25-inf)\''}
@attribute numVisitTheme4 {'\'(-inf-13]\'','\'(13-25]\'','\'(25-inf)\''}
@attribute effectivExeTheme4 {'\'(-inf-66.666667]\'','\'(66.666667-83.333333]\'','\'(83.333333-inf)\''}
@attribute effectivFINAL {'\'(-inf-75.646667]\'','\'(75.646667-87.253333]\'','\'(87.253333-inf)\''}

@data
101,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(-inf-20]\'','\'(16.666667-28.333333]\'','\'(20-35]\'','\'(15.33
102,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(-inf-20]\'','\'(16.666667-28.333333]\'','\'(-inf-20]\'','\'(15.
103,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(-inf-20]\'','\'(-inf-16.666667]\'','\'(-inf-20]\'','\'(-inf-15.
104,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(-inf-20]\'','\'(16.666667-28.333333]\'','\'(-inf-20]\'','\'(-in
105,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(20-35]\'','\'(16.666667-28.333333]\'','\'(35-inf)\'','\'(15.333
106,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(20-35]\'','\'(16.666667-28.333333]\'','\'(35-inf)\'','\'(15.333
107,'\'(-inf-68.333333]\'','\'(-inf-20]\'','\'(20-35]\'','\'(16.666667-28.333333]\'','\'(20-35]\'','\'(-inf-15
```

Figura 14. Archivo .arff con los datos discretizados

En este caso los atributos fueron etiquetados por tres rangos que luego se sustituyeron por las etiquetas, alta, media y baja.

Por ejemplo:

Para el atributo efectividad final a partir de los datos que se cuentan, el filtro retorna los rangos

{(-inf-75.646667], (75.646667-87.253333], (87.253333-inf)} que luego son remplazado por,

{baja, media, alta} y los datos quedan de la siguiente manera:

```
@attribute studentID numeric
@attribute visitThemes {baja,media,alta}
@attribute visitImagery {baja,media,alta}
@attribute visitVideos {baja,media,alta}
@attribute visitGlossary {baja,media,alta}
@attribute visitGames {baja,media,alta}
@attribute numExeTheme {baja,media,alta}
@attribute effectivExeTheme1 {baja,media,alta}
@attribute numVisitTheme2 {baja,media,alta}
@attribute effectivExeTheme2 {baja,media,alta}
@attribute numVisitTheme3 {baja,media,alta}
@attribute effectivExeTheme3 {baja,media,alta}
@attribute numVisitTheme4 {baja,media,alta}
@attribute effectivExeTheme4 {baja,media,alta}
@attribute effectivFINAL {baja,media,alta}

@data
101,baja,baja,baja,media,media,media,alta,media,alta,alta,alta,media,media,media
102,baja,baja,baja,media,baja,media,alta,media,alta,alta,alta,media,media,alta
103,baja,baja,baja,baja,baja,baja,alta,media,alta,alta,alta,media,alta,alta
104,baja,baja,baja,media,baja,baja,alta,baja,alta,alta,alta,media,alta,alta
105,baja,baja,media,media,alta,media,alta,alta,alta,alta,alta,baja,alta,alta
106,baja,baja,media,media,alta,media,alta,media,alta,alta,alta,media,alta,alta
107,baja,baja,media,media,media,baja,baja,baja,baja,baja,media,media,alta,baja
```

Figura 15. Archivo .arff con los datos etiquetados

Como salida de esta actividad se obtienen los datos listos para ejecutar el subproceso de Minería de datos.

2.3.3 Minería de datos

En esta fase se aplican las técnicas de modelado o de MD seleccionadas teniendo en cuenta las características de los datos o vista minable previamente seleccionada.

Fase 4: Minería de Datos

Las actividades realizadas en este subproceso se representan en la **Figura 16**, en esta fase se llevan a cabo dos actividades que se ejecutan indistintamente o una o la otra sin relaciones o dependencias entre ellas (ver **ANEXO 2**). A continuación se explica en detalles la ejecución de estas actividades.

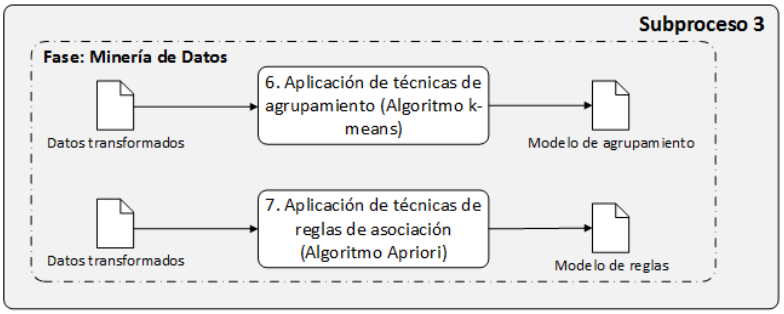


Figura 16. Descripción del Subproceso 3 [Elaboración propia]

Actividad 5: Aplicación de técnicas de agrupamiento

Para la tarea de agrupamiento fue seleccionado el algoritmo K-means. Se trata de un algoritmo clasificado como Método de Particionado y Recolocación (Garre et al., 2007). Los algoritmos de agrupamiento basados en particiones intentan buscar una división del conjunto de datos en subconjuntos con intersección vacía. Todos ellos siguen un patrón común que consiste en realizar una asignación de los objetos a los diferentes *clusters* en función de la proximidad de dichos objetos a un representante elegido para cada *cluster* (Lara, 2011).

El método de las k-medias es uno de los más utilizado en aplicaciones científicas e industriales. El nombre está dado porque representa cada uno de los *clusters* por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método únicamente se puede aplicar a atributos numéricos. Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. La suma de las discrepancias entre un punto y su centroide, expresado a través de la distancia apropiada, se usa como función objetivo. La función objetivo, suma de los cuadrados de los errores entre los puntos y sus centroides respectivos, es igual a la varianza total dentro del propio *cluster* (Hartigan, 1975; Hartigan & Wong, 1979).

En resumen el objetivo del algoritmo es dividir el conjunto *X* de los objetos en un cierto número *K* de subconjuntos naturales y homogéneos, donde los elementos de cada conjunto son tan similares como sea posible entre ellos y que, al mismo tiempo, sean lo más distinto posible a los demás integrantes de *X* (Pérez & León, 2007).

En la herramienta Weka se encuentra con el nombre de SimpleKMeans y en la solución propuesta se ejecuta con la siguiente configuración de los parámetros:

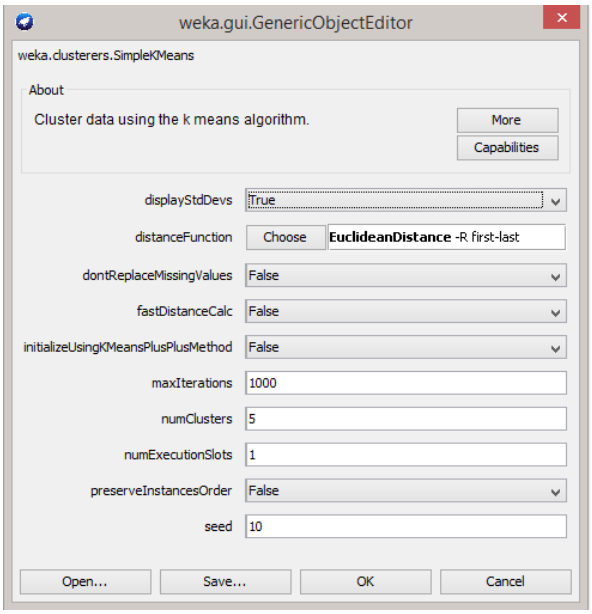


Figura 17. Configuración del algoritmo SimpleKmeans en la herramienta Weka

displayStdDevs: Para mostrar las desviaciones de los atributos numéricos y la cantidad de atributos nominales.

distanceFunction: Este parámetro selecciona la función de distancia, por lo que se regirá la semejanza/desemejanza de los objetos.

Distancia euclídea es la distancia ordinaria entre dos puntos calculada en un espacio euclídeo y se calcula a partir del teorema de Pitágoras. Dado dos puntos *A* y *B* medidos según las variables *X* y *Y* la distancia euclidiana seria (Ponce & Alcaraz, 2013; Soto & Jiménes, 2011):

$$d_{A-B} = \sqrt{(A_X - B_X)^2 + (A_y - B_y)^2} \qquad (1)$$

Cuando *A* y *B* estén medidas con un número *n* de dimensiones y no solo *X* y *Y* la fórmula sería la siguiente (Martín & Arias, 2015):

$$d_{A-B} = \sqrt{\sum_1^n (A_n - B_n)^2} \qquad (2)$$

dontRemplaceMissingValues: Se colocó en false debido a que no es necesario sustituir los valores faltantes por la media porque los datos omisos son tratados en la etapa de pre-procesamiento de datos.

numClusters: Cantidad de clúster a generar.

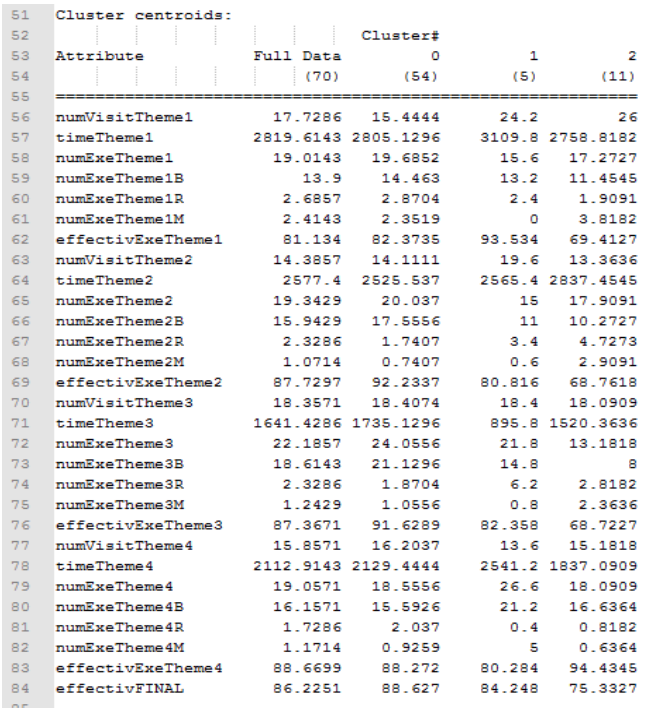
maxIterations: Número máximo de iteraciones, se estableció un numero alto para aumentar la fiabilidad del método.

Los demás parámetros se mantienen los valores que propone Weka por defecto. Para esta configuración el comando a ejecutar es el siguiente:

```
weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I
1000 -num-slots 1 -S 10
```

Una vez ejecutado este comando se obtiene un modelo como el de la **figura 18**, donde se muestra 3 grupos obtenidos con los valores de sus centroides (promedios para atributos numéricos, y valores más repetidos en cada grupo para atributos simbólicos) a partir de un conjunto de datos de prueba con 30 atributos y 70 instancias.
Las instancias por grupos quedaron repartidas de la siguiente manera:

Cluster	Cantidad de Instancias	Porciento
0	54	77 %
1	5	7%
2	11	16 %



```
Cluster centroids:
                                  Cluster#
Attribute            Full Data          0          1          2
                        (70)       (54)       (5)       (11)
=============================================================
numVisitTheme1         17.7286    15.4444       24.2         26
timeTheme1           2819.6143 2805.1296    3109.8 2758.8182
numExeTheme1           19.0143    19.6852       15.6    17.2727
numExeTheme1B             13.9     14.463       13.2    11.4545
numExeTheme1R           2.6857     2.8704        2.4     1.9091
numExeTheme1M           2.4143     2.3519          0     3.8182
effectivExeTheme1       81.134    82.3735     93.534    69.4127
numVisitTheme2         14.3857    14.1111       19.6    13.3636
timeTheme2              2577.4   2525.537     2565.4 2837.4545
numExeTheme2           19.3429      20.037         15    17.9091
numExeTheme2B          15.9429    17.5556         11    10.2727
numExeTheme2R           2.3286     1.7407        3.4     4.7273
numExeTheme2M           1.0714     0.7407        0.6     2.9091
effectivExeTheme2       87.7297    92.2337     80.816    68.7618
numVisitTheme3         18.3571    18.4074       18.4    18.0909
timeTheme3           1641.4286 1735.1296     895.8 1520.3636
numExeTheme3           22.1857    24.0556       21.8    13.1818
numExeTheme3B          18.6143    21.1296       14.8          8
numExeTheme3R           2.3286     1.8704        6.2     2.8182
numExeTheme3M           1.2429     1.0556        0.8     2.3636
effectivExeTheme3       87.3671    91.6289     82.358    68.7227
numVisitTheme4         15.8571    16.2037       13.6    15.1818
timeTheme4           2112.9143 2129.4444     2541.2 1837.0909
numExeTheme4           19.0571    18.5556       26.6    18.0909
numExeTheme4B          16.1571    15.5926       21.2    16.6364
numExeTheme4R           1.7286      2.037        0.4     0.8182
numExeTheme4M           1.1714     0.9259          5     0.6364
effectivExeTheme4       88.6699     88.272     80.284    94.4345
effectivFINAL           86.2251     88.627     84.248    75.3327
```

Figura 18. Ejemplo de modelo obtenido por el algoritmo K-means

Actividad 6: Aplicación de técnicas de reglas de asociación

Para la tarea de extraer reglas de asociación se utilizó el algoritmo Apriori. Este es uno de los algoritmos de reglas más utilizados. Se basa en la búsqueda de los conjuntos de ítems con determinada cobertura. Para ello, en primer lugar se construyen simplemente los conjuntos formados por sólo un ítem que supera la cobertura mínima. Este conjunto de conjuntos se utiliza para construir un conjunto de dos ítems, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan conjuntos de ítems con la cobertura requerida (Hernández-Orallo et al., 2004).

La propiedad plantea que todo subconjunto no vacío de un conjunto de ítems frecuentes tiene que ser frecuente también. Si un conjunto de ítems *X* no satisface el soporte mínimo entonces no es frecuente y si a este conjunto *X* se le adiciona un ítem entonces el conjunto resultante ocurrida con menor o igual frecuencia que *X* (Tanna & Ghodasara, 2014; I. Witten & E. Frank, 2005).

Para el uso de este a través de la herramienta Weka se necesita configurar los parámetros como se muestran en la **figura 19**.

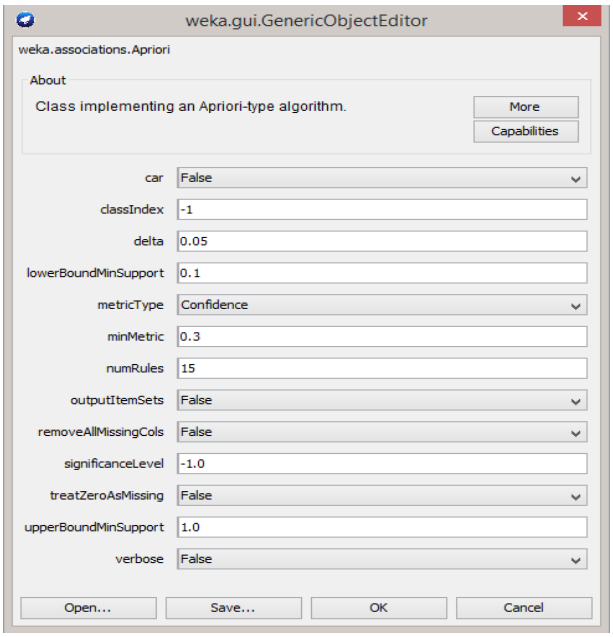


Figura 19.Configuración del algoritmo Apriori en la herramienta Weka

Características de los principales parámetros:

***car*:** Garantizar que se tengan en cuenta todos los atributos para propiciar un mayor número de reglas que si se tomase un solo atributo como clase.

classIndex*:** Se mantiene el valor por defecto ya que representa que se va a tomar como clase el último atributo, en caso de que se seleccione un solo atributo como clase en el parámetro ***car, en este se especifica qué atributo es el que se va a tomar.

lowerBoundMinSopport: Establece el soporte mínimo que la regla debe cumplir, en este caso se estableció el valor mínimo posible.

delta: Representa el valor que iterativamente va decreciendo el valor del soporte máximos hasta llegar al soporte mínimo.

metricType: Determina qué tipo de métrica se va a utilizar para determinar el porcentaje de calidad de las reglas.

minMetric: Mediante este parámetro se introduce la confianza mínima que las reglas a obtener tengan que cumplir. Este valor debe ser pequeño para poder obtener reglas que no se descubran a simple vista y así obtener información interesante.

numRules: Indica el número máximo de reglas a obtener, es decir, se utiliza como criterio de parada para detener la ejecución si se llega a este número de reglas cumpliendo las restricciones anteriores.

upperBoundMinSopport: Determina el soporte máximo a partir del cual se va a empezar a decrecer hasta llegar al soporte mínimo establecido, para el cual se definió el máximo posible.

Para esta configuración el comando a ejecutar queda de la siguiente forma:

```
weka.associations.Apriori -N 15 -T 1 -C 0.3 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
```

Una vez ejecutado este comando se obtiene un modelo como el de la **figura 20**, donde se muestra un fragmento con 15 de las reglas obtenidas a partir de un conjunto de datos de pruebas.

```
Best rules found:

 1. visitThemes=alta 18 ==> effectivFINAL=alta 14    conf:(0.78) < lift:(2.87)> lev:(0.13) [9] conv:(2.62)
 2. effectivFINAL=alta 19 ==> visitThemes=alta 14    conf:(0.74) < lift:(2.87)> lev:(0.13) [9] conv:(2.35)
 3. visitImagery=alta 20 ==> effectivFINAL=alta 14    conf:(0.7) < lift:(2.58)> lev:(0.12) [8] conv:(2.08)
 4. effectivFINAL=alta 19 ==> visitImagery=alta 14    conf:(0.74) < lift:(2.58)> lev:(0.12) [8] conv:(2.26)
 5. visitImagery=media 23 ==> numExeTheme=media effectivFINAL=media 15    conf:(0.65) < lift:(2.08)> lev:(0.11) [7] conv:(1.75)
 6. numExeTheme=media effectivFINAL=media 22 ==> visitImagery=media 15    conf:(0.68) < lift:(2.08)> lev:(0.11) [7] conv:(1.85)
 7. visitThemes=media 19 ==> effectivFINAL=media 15    conf:(0.79) < lift:(1.97)> lev:(0.11) [7] conv:(2.28)
 8. effectivFINAL=media 28 ==> visitThemes=media 15    conf:(0.54) < lift:(1.97)> lev:(0.11) [7] conv:(1.46)
 9. visitGlossary=media numExeTheme=media 20 ==> effectivFINAL=media 15    conf:(0.75) < lift:(1.88)> lev:(0.1) [6] conv:(2)
10. effectivFINAL=media 28 ==> visitGlossary=media numExeTheme=media 15    conf:(0.54) < lift:(1.88)> lev:(0.1) [6] conv:(1.43)
11. visitImagery=baja 27 ==> effectivFINAL=baja 16    conf:(0.59) < lift:(1.8)> lev:(0.1) [7] conv:(1.51)
12. effectivFINAL=baja 23 ==> visitImagery=baja 16    conf:(0.7) < lift:(1.8)> lev:(0.1) [7] conv:(1.77)
13. visitImagery=media numExeTheme=media 21 ==> effectivFINAL=media 15    conf:(0.71) < lift:(1.79)> lev:(0.09) [6] conv:(1.8)
14. effectivFINAL=media 28 ==> visitImagery=media numExeTheme=media 15    conf:(0.54) < lift:(1.79)> lev:(0.09) [6] conv:(1.4)
15. visitThemes=baja 33 ==> effectivFINAL=baja 19    conf:(0.58) < lift:(1.75)> lev:(0.12) [8] conv:(1.48)
```

Figura 20. Ejemplo de modelo obtenido por el algoritmo Apriori

Para cada regla, se muestran las frecuencias de ocurrencia (cobertura) para el lado izquierdo y lado derecho de cada regla, así como los valores de la confianza (conf), la elevación (lift), el apalancamiento (lev), y la convicción (conv).

Estos modelos creados por Weka ya sea el de Agrupamiento o el de Reglas es salvado en un archivo ARFF que constituye la salidad de esta actividad.

2.3.4 Análisis

Al llegar a esta fase se analizarán los modelos seleccionados para determinar si cumplen apropiadamente con los objetivos del negocio propuestos en la primera **fase**.

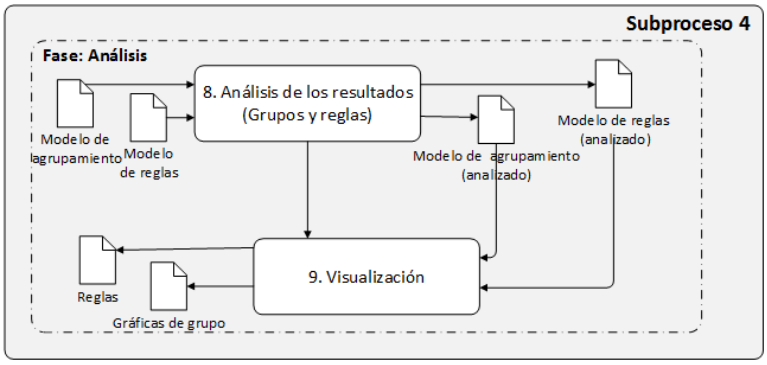


Figura 21. Subproceso 4 propuesto [Elaboración propia]

Actividad 7: Análisis de los resultados

Para lograr esto la evaluación se apoya de dos sub-actividades: la evaluación de los resultados (evaluación de los resultados de la aplicación de la minería de datos, modelos aprobados) y la revisión del proceso.

Evaluación de los resultados

Como se planteaba en la fundamentación teórica de esta investigación medir la calidad de los patrones descubiertos por un algoritmo de MD no es un problema trivial y aún más complicado cuando se trata de modelos descriptivos, donde no existe una clase determinada o un valor numérico donde medir el grado de acierto del modelo obtenido. Es por ello que las medidas de evaluación de modelos descriptivos se basan en conceptos tales como la complejidad del modelo y de los datos a partir del modelo, o bien, en agrupamiento, el nivel de compactación de los diferentes grupos.

Al no contar con datos reales para probar el proceso, se hace aún más difícil establecer una evaluación de los modelos obtenidos, por lo que se contará principalmente con la opinión de posibles usuarios de la aplicación al mostrarles cuales pueden ser los resultados. Teniendo en cuenta criterios como:

- El interés que pueda despertar en los usuarios los modelos obtenidos.
- La capacidad de un modelo de sorprender a los usuarios con respecto al conocimiento previo que tenía sobre determinado problema (novedad).
- La comprensibilidad o inteligibilidad de los resultados obtenidos.
- El tamaño o complejidad del modelo.
- La capacidad de ser utilizado con éxito en el contexto real donde va a ser aplicado (Aplicabilidad).

Esto se puede ver con más detalles en el **Capítulo 3**.

Revisión del proceso

Con la fase de la comprensión del negocio, que constituye su etapa inicial, centrada en entender los objetivos y metas fundamentales del negocio, se establece como objetivo fundamental del negocio, generar un reporte que sea capaz de exponer de forma clara y sencilla elementos importantes y de relevancia almacenados en el módulo Resultados de la plataforma educativa Navigo.

Una vez determinado los objetivos se dio paso a una de las fases más importantes del proceso, la selección de los datos, que por su relevancia se deben tener en cuenta a la hora de realizar el proceso; el acceso de los estudiantes a los diferentes módulos de la colección y los resultados obtenidos en las evaluaciones del módulo Ejercicios.

Una vez seleccionados los datos se procedió a la preparación de estos para poder ser interpretados correctamente por la herramienta de MD, transformándolos siguiendo la estructura del fichero de extensión ARFF que propone Weka.

En la siguiente fase se aplicaron las tareas y algoritmos de MD seleccionados teniendo en cuenta las características de los datos que se eligieron. Se decide la realización de tareas basadas en agrupamientos y reglas de asociación. Se seleccionaron como algoritmos para la aplicación de estas tareas, relacionados con el agrupamiento a K-means y por parte de las reglas de asociación Apriori.

Actividad 8: Visualización

Las tecnologías de la visualización muestran gráficamente los datos en las bases de datos. Se ha investigado mucho sobre la visualización y el campo ha adelantado un gran trecho sobre todo con la incorporación de la informática multimedia. Por ejemplo, los datos en las bases de datos serán filas y filas de valores numéricos, y las herramientas de visualización toman estos datos y trazan con ellos algún tipo de gráfico. Los modelos de visualización pueden ser bidimensionales, tridimensionales o incluso multidimensionales. Se han desarrollado varias herramientas de visualización para integrarse con las bases de datos, y algunos trabajos sobre este tema están recogidos en (U. M. Fayyad, Wierse, & Grinstein, 2002).

Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos (Hasperué, 2012):

- Aprovechar la gran capacidad humana de ver patrones, anomalías y tendencias a partir de imágenes y facilitar la compresión de los datos.
- Ayudar al usuario a comprender rápidamente patrones descubiertos automáticamente por un sistema de extracción de conocimiento.

La entrada a esta actividad son los modelos expuestos en la **Fase 4**. Estos son archivos ARFF que la plataforma educativa Navigo se encarga de interpretarlos (ver **epígrafe 2.4**) y a partir de la información que contienen estos modelos, son construidos gráficos que reflejan los distintos grupos que resultaron de la aplicación de los algoritmos de agrupamiento y también las reglas generadas por las tareas de reglas de asociación, todo esto de la forma más sencilla posible como se muestra en las siguientes imágenes.

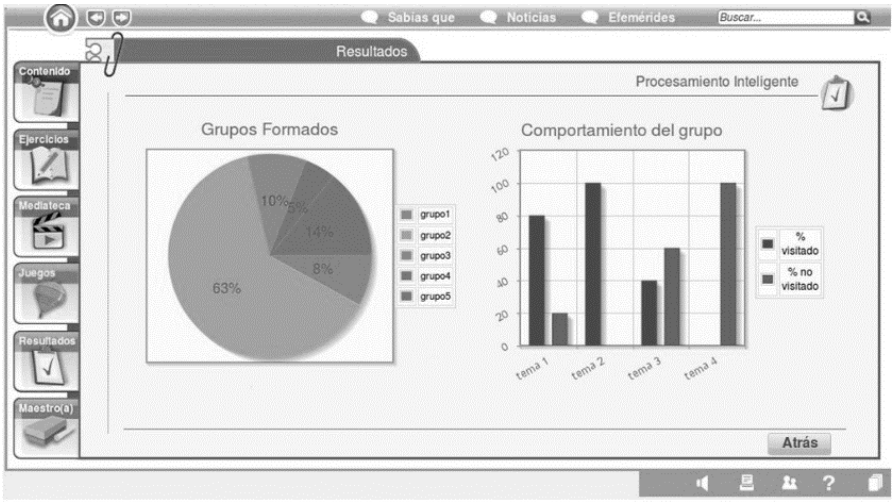


Figura 22. Visualización del modelo de agrupamiento

Figura 23. Visualización de las reglas de asociación

2.4 Integración de Weka con el hiperentorno

Como se plantea en la problemática de esta investigación unos de los principales retos que tiene la comunidad de EDM es la integración de algoritmos y herramientas de MD con los diferentes tipos de entornos o plataformas educativas, como también lo es el desarrollo de herramientas fáciles e intuitivas de utilizar para los docentes, estudiantes o cualquier usuario sin la necesidad de tener conocimientos sobre las técnicas de MD y su uso.

Para desarrollar el proceso de KDD desde la plataforma educativa Navigo se decidió integración de esta con una herramienta que apoye el ejercicio de MD, permitiendo la reutilización de técnicas y algoritmos ya implementados. Como se explica en el **epígrafe 1.8** la herramienta seleccionada fue Weka y se tuvieron en cuenta los retos anteriormente mencionados.

2.4.1 Características arquitectónicas de la plataforma educativa Navigo

La plataforma educativa Navigo está desarrollada bajo la arquitectura Cliente-Servidor de una aplicación web, siguiendo el patrón arquitectónico Modelo-Vista-Controlador (MVC) que propone el *framework* Symfony sobre el cual se desarrolló la aplicación.

Las tecnologías y lenguajes que se utilizaron para el desarrollo de la plataforma son:

- HTML, CSS, JavaScript y PHP
- *Framework* Symfony v1.4.3
- ORM Doctrine
- JQuery v1.5.1 y JQueryUI v1.8
- Servidor de Aplicaciones Apache v2.0
- Sistema Gestor de Base de Datos PostgresSQL v9.4

Como es una aplicación diseñada para la web y desarrollada con tecnologías libres tiene el propósito de no tener ninguna restricción en cuanto a sistema operativo ya sea libre o propietario. Estas características influenciaron en la selección de la herramienta de MD para realizar el proceso de KDD propuesto.

2.4.2 Integración de la solución propuesta con la plataforma educativa Navigo

Después de estudiar varias opciones se decidió utilizar como vía de comunicación entre Navigo y Weka, a través de líneas de comandos. Esta es una funcionalidad que implementa Weka y permite entonces la comunicación a través de la invocación de un comando desde la aplicación web. Para esto se decidió utilizar la función `system` de php que permite la ejecución de un programa externo y obtener su salida.

La sintaxis de la función es:

```
string system ( string $command [, int &$return_var ] )
```

Los parámetros:

`$command`: El comando que será ejecutado.

`$return_var`: Si el argumento `return_var` se encuentra presente, entonces el status devuelto por el comando ejecutado será almacenado en esta variable.

El comando para ejecutar el archivo Weka.jar sería:

```
system("java -Xmx [MEGABYTES_DE_MEM_PARA_LA_TAREA] -cp [PATH_A_weka.jar]
[ALGORITMO+PARÁMETROS][FICHERO_DATOS_ARFF]> [MODELO_SALIDA]", $return_var);
```

Teniendo en cuenta las configuraciones de los algoritmos expuestas en la **actividad 5** del proceso propuesto, para la tarea de Agrupamiento, la ejecución de K-means quedaría de la siguiente forma:

```
system("java -Xmx1024M -cp weka/weka.jar  weka.clusterers.SimpleKMeans -V -N
"5" -A \"weka.core.EuclideanDistance -R first-last\" -I 1000 -S 10 -t
weka/dataNAVIGO/datosA.arff > weka/dataNAVIGO/modeloCluster.arff", $retval);
```

Para las reglas de asociación sería:

```
system("java -Xmx512M -cp weka/weka.jar  weka.associations.Apriori -N 10 -T 0
-C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1 -t weka/dataNAVIGO/datosR.arff >
weka/dataNAVIGO/modeloRule.arff", $retval);
```

Para un mayor entendimiento de cómo se integra la solución implementada con la plataforma educativa Navigo, en la **figura 24** se hace una descripción a través de un diagrama donde se muestra el flujo o secuencia de actividades llevadas a cabo en la integración de la plataforma y la herramienta Weka en la solución implementada.

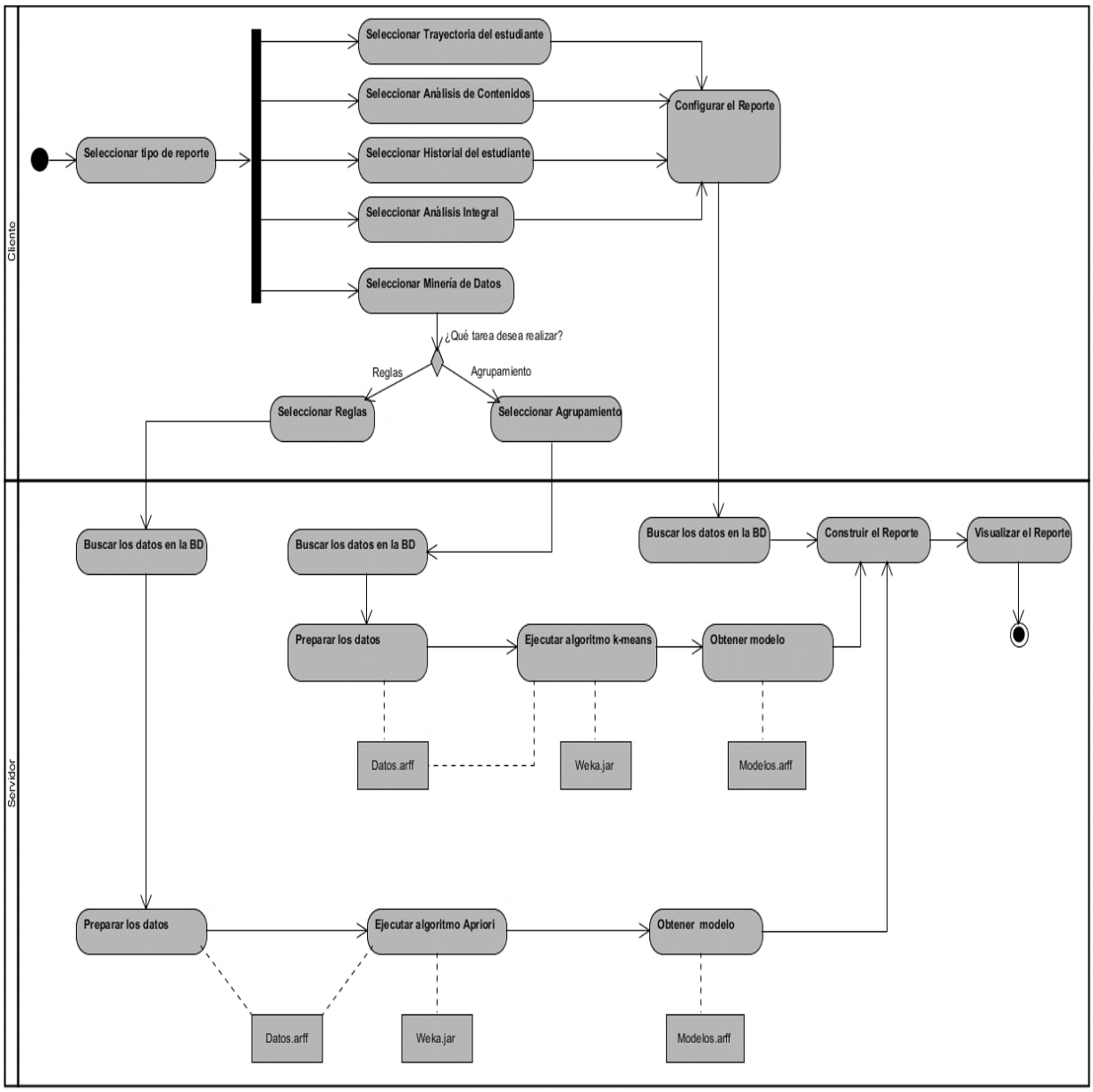


Figura 24. Secuencia de actividades en la integración de la solución

En la **figura 25** se describe técnicamente las relaciones y dependencias de clases y componentes de la plataforma Navigo y Weka.

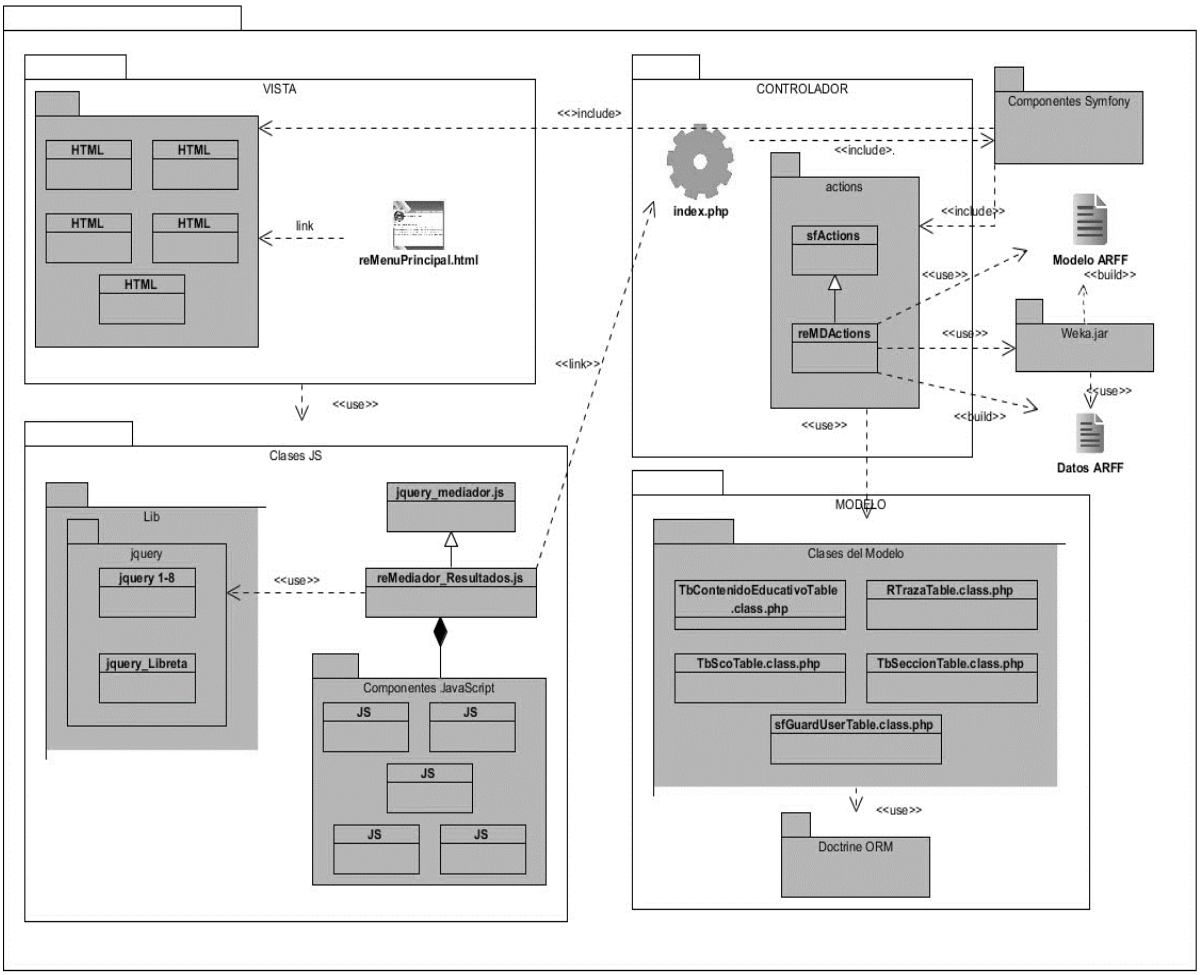


Figura 25. Modelado de la integración a nivel de diseño

Conclusiones del capítulo

Una vez descrita la solución implementada, se relacionan a continuación las consideraciones finales del presente capítulo:

- La implementación de las funcionalidades, Trayectoria del estudiante, Análisis de contenidos, Historial del estudiante y Análisis constituyen un aporte a la concepción del módulo Resultados de la plataforma educativa Navigo, que resuelve las deficiencias existentes en las versiones anteriores de hiperentornos.
- Se definió un proceso de KDD integrado por cuatro subprocesos: Definición de los objetivos del proceso; Preparación de los datos; Minería de datos, donde se aplican las tareas de agrupamiento y reglas de asociación mediante los algoritmos K-means y Apriori respectivamente y el subproceso de Análisis para descubrir grupos de estudiantes con comportamientos similares, patrones de navegación o posibles causas de las evaluaciones
- La integración de la plataforma con Weka como herramienta de apoyo para la realización de las tareas de MD facilitó la implementación y ejecución del proceso de KDD desde la propia plataforma.
- Con la incorporación de técnicas de MD y reportes estadísticos, la plataforma cuenta con una herramienta que contribuyen al control y seguimiento del aprendizaje de los alumnos a partir de la interacción de estos con los hiperentornos.

CAPÍTULO 3
VALIDACIÓN DE LA PROPUESTA

Lo oí y lo olvidé. Lo vi y lo entendí. Lo hice y lo aprendí.
CONFUCIO

En el presente capítulo se describe la validación de los resultados de la investigación y se explican los métodos utilizados. Para validar la propuesta se realizaron pruebas funcionales a la plataforma, luego de ello se aplicaron otras técnicas como la escala de Likert fundamentada en la opinión de los expertos para validar los modelos obtenidos y la solución propuesta.

3.1 Pruebas de liberación

Como en todo desarrollo de software una vez terminada la fase de implementación de la plataforma educativa Navigo se comenzó a ejecutar la fase de Pruebas. Primeramente se llevaron a cabo pruebas internas por el equipo del proyecto, en una segunda etapa se procedió a las pruebas por el Grupo de Calidad de FORTES y finalmente se pasó a la fase de liberación por el laboratorio de pruebas de la dirección de Calidad de la UCI. En la etapa de Planificación de la estrategia se pactó la realización de pruebas funcionales a cada uno de los módulos de la plataforma para comprobar si los mismos cumplían las funciones esperadas. Fue necesario para que se desarrollaran dichas pruebas entregarle a Calidad UCI, como complemento de la plataforma, la Especificación de Requisitos y todos los Casos de Pruebas previamente diseñados (tomando como referencia las especificaciones funcionales).

La etapa de ejecución de las pruebas se desarrolló en 3 iteraciones y una prueba final donde fueron detectadas y resueltas un conjunto de No Conformidades (NC) como se muestra en la tabla 4 (Casañola, 2014).

Tabla 4. Resultados de las pruebas de liberación por Calidad-UCI [Tomado de(CALIDAD-UCI, 2014)]

Iteración	No conformidades	
	Detectadas	Resueltas
1ra	60	77% (el resto no procedieron)
2da	40	85% (el resto no procedieron)
3ra	25	100%
PF	0	---

De estas NC, 10 fueron detectadas en módulo de Resultados, de ellas el 100% fueron resueltas. La mayoría de estas NC no estaban relacionadas con la solución propuesta en esta investigación, aunque al respecto se detectaron 3 NC relacionadas con el registro de las trazas lo cual si influye directamente en los requisitos relacionados con la presente investigación. Las 3 NC fueron de funcionalidad, para las cuales se establecieron las acciones correctivas necesarias para llegar a la prueba final sin ninguna NC pendiente y obtener el acta de liberación de la plataforma educativa Navigo.

Con la resolución de estas NC quedó validado el cumplimiento de cada uno de los requisitos funcionales para la implementación de técnicas de MD como parte fundamental de un proceso de KDD, así como un grupo de reportes estadísticos que aportaron nuevas funcionalidades didácticas

al módulo Resultados de la plataforma educativa Navigo. Estas funcionalidades contribuyen al control y seguimiento personalizado del aprendizaje de cada estudiante o grupo en particular a partir de los datos almacenados por la plataforma.

3.2 Validación de los modelos descriptivos

En el momento que se realiza esta investigación la plataforma educativa Navigo no se encuentra desplegada en ningún centro educacional por lo que no se cuentan con registros reales almacenados en su BD. Esto dificulta las pruebas y validaciones sobre todas las funcionalidades presentes en el módulo Resultados. Para enmendar esta dificultad se publicó la plataforma en un servidor de pruebas y se crearon los datos de una escuela *X*, dos grupos académicos por cada año (séptimo, octavo y noveno) y diez usuarios como estudiantes por cada grupo, para un total de 60 estudiantes de la escuela *X*.

Con el objetivo de obtener un número de registros para luego ser procesados y mostrar reportes, en un ambiente controlado se repartieron los usuarios de los estudiantes creados a los miembros del equipo de desarrollo de la plataforma y varios colaboradores (especialistas del centro FORTES) para que accedieran a la plataforma e interactuaran con los diferentes recursos presentes en un hiperentorno educativo, simulando de esta forma el uso de la plataforma en un entorno real. Por otra parte se cuenta con las trazas obtenidas durante las fases de pruebas por el Grupo de Calidad de FORTES y el laboratorio de pruebas de Calidad-UCI. También se aprovecharon las diferentes ferias en que se ha expuesto la plataforma para atraer usuarios en edades de secundaria básica y obtener los registros de su interacción con la plataforma.

Las técnicas de MD están diseñadas para trabajar con grandes volúmenes de datos pero al no estar en explotación la plataforma educativa Navigo, solo se cuentan con el registro de 397 sesiones de trabajo para probar la solución propuesta.

Después de ejecutar el aloritmo Apriori (para esta configuración **weka.associations.Apriori -N 50 -T 0 -C 0.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c -1**) se obtuvo el siguiente conjunto de reglas:

```
 1. visitThemes=media visitVideos=media 15 ==> effectivFINAL=media 15    conf:(1)
 2. visitThemes=media visitImagery=media 13 ==> effectivFINAL=media 13    conf:(1)
 3. visitThemes=media visitImagery=media numExeTheme=media 13 ==> effectivFINAL=media 13
conf:(1)
 4. visitThemes=alta visitImagery=alta 12 ==> effectivFINAL=alta 12    conf:(1)
 5. visitThemes=alta numExeTheme=alta 12 ==> effectivFINAL=alta 12    conf:(1)
 6. visitImagery=alta numExeTheme=alta 11 ==> effectivFINAL=alta 11    conf:(1)
 7. visitThemes=media visitVideos=media numExeTheme=media 11 ==> effectivFINAL=media 11
conf:(1)
 8. visitThemes=media visitGames=media numExeTheme=media 11 ==> effectivFINAL=media 11
conf:(1)
 9. visitThemes=alta visitImagery=alta numExeTheme=alta 10 ==> effectivFINAL=alta 10
conf:(1)
10. visitImagery=alta visitVideos=alta 9 ==> effectivFINAL=alta 9    conf:(1)
11. visitGames=alta numExeTheme=alta 9 ==> effectivFINAL=alta 9    conf:(1)
12. visitThemes=media visitImagery=media visitGlossary=media 9 ==> effectivFINAL=media 9
conf:(1)
13. visitThemes=alta visitImagery=alta visitGames=alta 9 ==> effectivFINAL=alta 9    conf:(1)
14. visitImagery=media visitVideos=media numExeTheme=media 9 ==> effectivFINAL=media 9
conf:(1)
15. visitGlossary=media visitGames=media numExeTheme=media 9 ==> effectivFINAL=media 9
conf:(1)
16. visitThemes=media visitImagery=media visitGlossary=media numExeTheme=media 9 ==>
effectivFINAL=media 9    conf:(1)
17. visitThemes=media visitVideos=media visitGlossary=media 8 ==> effectivFINAL=media 8
conf:(1)
18. visitThemes=alta visitGlossary=media numExeTheme=alta 8 ==> effectivFINAL=alta 8
conf:(1)
19. visitThemes=alta visitGames=alta numExeTheme=alta 8 ==> effectivFINAL=alta 8    conf:(1)
20. visitImagery=alta visitGames=alta numExeTheme=alta 8 ==> effectivFINAL=alta 8    conf:(1)
```

```
21. visitThemes=media visitImagery=media visitVideos=media 7 ==> effectivFINAL=media 7
conf:(1)
22. visitThemes=alta visitImagery=alta visitGlossary=media 7 ==> effectivFINAL=alta 7
conf:(1)
23. visitImagery=alta visitGlossary=media numExeTheme=alta 7 ==> effectivFINAL=alta 7
conf:(1)
24. visitVideos=media visitGlossary=media numExeTheme=media 7 ==> effectivFINAL=media 7
conf:(1)
25. visitThemes=media visitImagery=media visitVideos=media numExeTheme=media 7 ==>
effectivFINAL=media 7    conf:(1)
26. visitThemes=media visitGlossary=media visitGames=media numExeTheme=media 7 ==>
effectivFINAL=media 7    conf:(1)
27. visitThemes=alta visitImagery=alta visitGlossary=media numExeTheme=alta 7 ==>
effectivFINAL=alta 7    conf:(1)
28. visitThemes=alta visitImagery=alta visitGames=alta numExeTheme=alta 7 ==>
effectivFINAL=alta 7    conf:(1)
29. visitGames=media numExeTheme=media 14 ==> effectivFINAL=media 13    conf:(0.93)
30. visitThemes=media visitGlossary=media numExeTheme=media 14 ==> effectivFINAL=media 13
conf:(0.93)
31. visitImagery=alta visitGames=alta 13 ==> effectivFINAL=alta 12    conf:(0.92)
32. visitImagery=media visitVideos=media 10 ==> effectivFINAL=media 9    conf:(0.9)
33. visitGlossary=media numExeTheme=alta 9 ==> effectivFINAL=alta 8    conf:(0.89)
34. visitImagery=media visitGames=media 8 ==> effectivFINAL=media 7    conf:(0.88)
35. visitVideos=alta visitGames=alta 8 ==> effectivFINAL=alta 7    conf:(0.88)
36. visitThemes=media visitGlossary=media visitGames=media 8 ==> effectivFINAL=media 7
conf:(0.88)
37. visitImagery=media visitGames=media numExeTheme=media 8 ==> effectivFINAL=media 7
conf:(0.88)
38. visitVideos=baja visitGames=media numExeTheme=media 8 ==> effectivFINAL=media 7
conf:(0.88)
39. visitGlossary=media numExeTheme=media 21 ==> effectivFINAL=media 18    conf:(0.86)
40. visitThemes=media visitGames=media 14 ==> effectivFINAL=media 12    conf:(0.86)
41. visitImagery=media visitGlossary=media numExeTheme=media 14 ==> effectivFINAL=media 12
conf:(0.86)
42. visitThemes=alta visitGames=alta 13 ==> effectivFINAL=alta 11    conf:(0.85)
43. visitThemes=media numExeTheme=media 24 ==> effectivFINAL=media 20    conf:(0.83)
44. visitImagery=baja visitGlossary=baja 12 ==> effectivFINAL=baja 10    conf:(0.83)
45. visitImagery=media numExeTheme=media 22 ==> effectivFINAL=media 18    conf:(0.82)
46. visitImagery=media visitGlossary=media 15 ==> effectivFINAL=media 12    conf:(0.8)
47. visitImagery=baja visitVideos=baja 10 ==> effectivFINAL=baja 8    conf:(0.8)
48. visitImagery=media visitGames=baja 10 ==> effectivFINAL=media 8    conf:(0.8)
49. visitVideos=baja visitGames=baja 10 ==> effectivFINAL=baja 8    conf:(0.8)
50. visitImagery=media visitGames=baja numExeTheme=media 10 ==> effectivFINAL=media 8
conf:(0.8)
```

Como se puede observar para el caso de las reglas de asociación, las reglas obtenidas tienen un valor de confianza por encima de 0.8, se pueden conocer algunas reglas interesantes aunque otras lo son menos. Por ejemplo varias reglas indican, como es de esperar, que los estudiantes que tienen una frecuencia alta de visitas al módulo Contenido y otros recursos como galerías y juegos por lo general tienen una efectividad final alta, esto se cumple casi que en un 100% según los registros procesados.

Para (Orallo, Quintana, & Ramírez, 2006) la calidad de las reglas de asociación como resultado del aprendizaje muchas veces viene lastrada por la presencia de atributos que estén fuertemente descompensados. Por ejemplo, en este caso la escasa presencia de frecuencia baja de visitas al módulo Contenido provoca que no aparezcan en las reglas de asociación, ya que las reglas con este *ítemset* poseen una baja cobertura y son filtradas.

Con respecto al agrupamiento el modelo obtenido se muestra a continuación:

		Cluster#		
Attribute	**Full Data**	**0**	**1**	**2**
	(97)	**(15)**	**(31)**	**(51)**
==========	==========	==========	==========	==========
numVisitTheme1	**17.4536**	**20.3333**	**18.871**	**15.7451**
timeTheme1	**2795.0206**	**2713.9333**	**2428.0645**	**3041.9216**
numExeTheme1	**18.9278**	**18.1333**	**20.6774**	**18.098**

numExeTheme1B	**13.3299**	**15.7333**	**10.8065**	**14.1569**
numExeTheme1R	**2.8763**	**1.8667**	**4.4516**	**2.2157**
numExeTheme1M	**2.7113**	**0.5333**	**5.3871**	**1.7255**
effectivExeTheme1	**79.0786**	**91.9253**	**61.84**	**85.7784**
numVisitTheme2	**14.567**	**21**	**13.6774**	**13.2157**
timeTheme2	**2571.1134**	**2478.8667**	**2241.4194**	**2798.6471**
numExeTheme2	**19.8247**	**18.6667**	**15.5161**	**22.7843**
numExeTheme2B	**16.4948**	**15.2**	**11.0645**	**20.1765**
numExeTheme2R	**2.3299**	**2.7333**	**2.5484**	**2.0784**
numExeTheme2M	**1**	**0.7333**	**1.9032**	**0.5294**
effectivExeTheme2	**88.2273**	**87.7433**	**81.3503**	**92.5498**
numVisitTheme3	**18.2784**	**19.0667**	**16.8387**	**18.9216**
timeTheme3	**1626.1959**	**1796.8667**	**1544.2581**	**1625.8039**
numExeTheme3	**22.2062**	**24.6**	**14.6452**	**26.098**
numExeTheme3B	**18.732**	**17.4**	**10.6774**	**24.0196**
numExeTheme3R	**2.2887**	**5.8**	**2.3871**	**1.1961**
numExeTheme3M	**1.1856**	**1.4**	**1.5806**	**0.8824**
effectivExeTheme3	**87.9659**	**82.2813**	**81.4303**	**93.6104**
numVisitTheme4	**15.9175**	**14.2**	**16.1613**	**16.2745**
timeTheme4	**2015.5567**	**2386.3333**	**1714.1613**	**2089.7059**
numExeTheme4	**19.1237**	**18.1333**	**18.8387**	**19.5882**
numExeTheme4B	**16.2062**	**14.2**	**16.6452**	**16.5294**
numExeTheme4R	**1.7526**	**0.8**	**1.7742**	**2.0196**
numExeTheme4M	**1.1649**	**3.1333**	**0.4194**	**1.0392**
effectivExeTheme4	**88.8923**	**76.1767**	**93.5613**	**89.7941**
effectivFINAL	**86.0374**	**84.532**	**79.5448**	**90.4267**

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 15 (15%)
1 31 (32%)
2 51 (53%)

Como se ha explicado en esta investigación los modelos descriptivos son difíciles de evaluar. Teniendo en cuenta lo que se plantea en (Hernández-Orallo et al., 2004), la mejor evaluación de este tipo de modelos es saber si el modelo obtenido de la fase de aprendizaje tiene un comportamiento útil cuando se utiliza en su área de aplicación. Este criterio está muy relacionado con el interés, la novedad, la aplicabilidad, así como la comprensibilidad y la simplicidad de los modelos. Por tanto se decide consultar el criterio de expertos para validar los posibles modelos a obtener según los criterios planteados.

Para obtener una evaluación según estos criterios se decidió aplicar una encuesta (ver **ANEXO 2**) a un grupo de profesores de la enseñanza secundaria. Para lograr este fin, resultaría muy difícil que el investigador pudiera trabajar con una muestra representativa de toda la población de escuelas secundarias del país. Por lo anterior, fueron seleccionadas intencionalmente la ESBU Rubén Martínez Villena y el ESBEC Andrés Cueva Heredia pertenecientes a la localidad de Vueltas donde reside el autor de la presente investigación, en la provincia de Villa Clara, donde además la dirección de ambas facilitaron el espacio y tiempo para llevar a cabo la encuesta. Para

esto se seleccionaron igualmente mediante un muestro intencionado 23 profesores que cumplieran con los requisitos de tener más de 10 años de experiencia vinculados a la docencias en la enseñanza y que tuvieran conocimiento y experiencia en el trabajo con la colección "El Navegante".

En el gráfico de la **figura 26** se describe la composición de la muestra de acuerdo a los años de experiencia en la educación y la frecuencia con que utilizan el software educativo como apoyo al proceso de enseñanza y aprendizaje.

La muestra tiene una media de 17 años de experiencia como docentes, el 35% es o ha sido profesor de computación, el 26% tienen la categoría científica de Master en Ciencias de la Educación y el 100 % conoce y ha utilizado la colección "El Navegante".

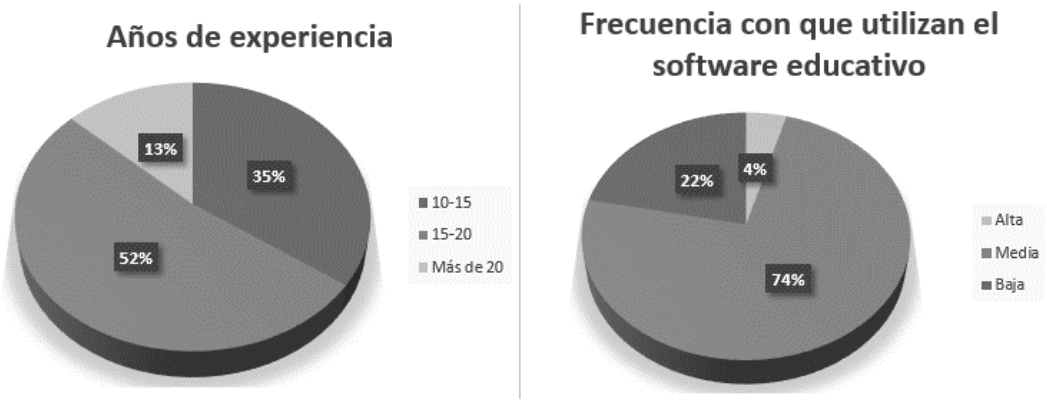


Figura 26. Descripción de la muestra [Elaboración propia]

Es importante destacar que el objetivo de esta encuesta es evaluar la validez y aplicabilidad de los modelos obtenidos en esta investigación y no extrapolar los resultados de estas dos escuelas al resto de las secundarias del país.

A partir de los datos recopilados por la encuesta (ver **ANEXO 4**) se calculó un índice porcentual (IP), que integra en un solo valor la evaluación de los docentes sobre los criterios propuestos. El cálculo se realiza según la **fórmula (4)**, adaptada a las categorías de evaluación establecidas en la encuesta donde uno significa poco valor y cinco significa el máximo valor posible. En la gráfica de la **figura 27** se muestran el IP por cada uno de los criterios evaluados.

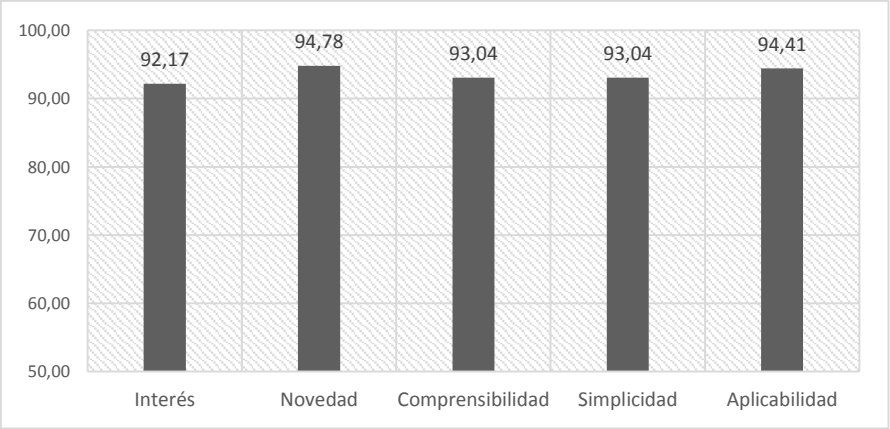


Figura 27. Resultados de la encuesta para validar los modelos

Como se puede observar todos los indicadores están por encima del 90 % lo cual es un resultado satisfactorio. Además del IP calculado por cada criterio, para conocer la medida de acuerdo entre los docentes encuestados se calculó el coeficiente de concordancia mediante la prueba *W de Kendall*. El procesamiento de los datos se realizó utilizando la herramienta SPSS y como salida de la prueba estadística se obtuvo:

Tabla 5. Salida de la prueba estadística W de Kendall

N	5
W de Kendall[a]	,539
Chi-cuadrado	20,755
gl	22
Sig. asintótica	,536

a. Coeficiente de concordancia de Kendall

Por tanto se rechaza la hipótesis de concordancia casual y se obtiene un coeficiente de concordancia mayor que cero demostrando que los expertos están de acuerdo en los criterios evaluados sobre los modelos de reglas y agrupamiento que se obtienen en la plataforma educativa Navigo.

3.3 Criterio de expertos sobre la propuesta

En ocasiones se cree que la aplicación de encuestas es exclusiva de las ciencias sociales, sin embargo, esta herramienta se ha difundido considerablemente en los últimos tiempos entre los especialistas e investigadores de diversas ramas de las ciencias técnicas. Para ello se parte de la premisa de que, si se quiere conocer algo sobre el comportamiento de determinado fenómeno, lo mejor y más simple, es preguntarle directamente a las personas que están relacionadas de una forma u otra con este y poseen la mayor experiencia en tema.

Entre los aspectos más importantes para obtener resultados confiables durante la aplicación de encuestas a expertos, se encuentra la determinación del número mínimo de candidatos que tomarían parte en la misma (grupo significativo de personas), así como los que pueden ser considerados como expertos. Teniendo en cuenta lo planteado en (L. García & Fernández, 2008) la cantidad óptima de expertos a consultar para la aplicación del método, oscila entre 15 y 25, este criterio es avalado también por la experiencia de diferentes autores en la actividad docente-investigativa y la aplicación de este método en investigaciones por más de 25 años por parte del autor del artículo referenciado.

En el caso de la presente investigación para la selección de los expertos no se utilizaron muestreos probabilísticos, pues el propósito no era buscar representatividad en el sentido estadístico, sino garantizar juicios autorizados en el tema y por consiguiente validez en la información. Para ello se utilizó el muestreo discrecional o también llamado intencional (Dios, 2015), que consiste en la selección de casos ricos en información, de los cuales se pueden extraer conclusiones de relevancia para los propósitos de la presente investigación.

La cantidad de expertos seleccionados fue de 18 especialistas que a criterio del autor cumplían los requisitos de tener una vasta experiencia como docentes en la enseñanza secundaria y en el trabajo con el software educativo dentro del proceso de enseñanza y aprendizaje o que fueran especialistas con experticia en el desarrollo de software educativo. A este grupo se les explicó detalladamente en que consiste la investigación, se les mostró el funcionamiento del módulo Resultados de la plataforma Navigo y luego se les aplicó un cuestionario (ver **ANEXO 3**) con el

objetivo de evaluar finalmente la propuesta que se hace en el presente trabajo. A este cuestionario se le agregó un Test de autoevaluación para caracterizar estadísticamente las competencias de cada uno.

A partir de los datos obtenidos en el Test se calcula el coeficiente de competencia (*K*) para lo que se tiene en cuenta la autovaloración del experto acerca de su competencia (*Kc* coeficiente de conocimiento) y las fuentes de argumentación (*Ka* coeficiente de argumentación) mediante la siguiente fórmula (3):

$$K = \frac{Kc+Ka}{2} \quad (3)$$

Donde:

Kc: es el coeficiente de conocimiento o información que tiene el experto acerca del problema, calculado sobre la valoración del propio experto en una escala del 1 al 10, de esta forma, la evaluación "1" indica que el experto tiene muy poco conocimiento de la problemática correspondiente, mientras que la evaluación "10" significa todo lo contrario. En la siguiente tabla se muestra la relación del valor de ***Kc*** por cada uno de los expertos.

Ka: es el coeficiente de argumentación o fundamentación de los criterios del experto, obtenido como resultado de la suma de los puntos alcanzados a partir de una tabla patrón como la siguiente:

Tabla 6. Tabla para medir el grado de influencia de las fuentes de argumentación

Fuentes de argumentación	Grado de influencia de cada una de las fuentes en sus criterios.		
	Alto	Medio	Bajo
Análisis teóricos realizados.	0.3	0.2	0.1
Experiencia obtenida.	0.5	0.4	0.3
Trabajos de autores nacionales.	0.05	0.04	0.03
Trabajos de autores extranjeros.	0.05	0.04	0.03
Su propio conocimiento del tema.	0.05	0.04	0.03
Intuición.	0.05	0.04	0.03

Después del cálculo de ***K***, se determina el nivel de competencia según la siguiente tabla:

Tabla 7. Escala para medir grado de competencia

Coeficiente de competencia	Valor
Alto	0,8 < K < 1,0
Medio	0,5 < K < 0,8
Bajo	K < 0,5

En la tabla 8 se muestra un resumen de los grados y coeficiente de competencia, coeficiente de conocimiento y coeficiente de argumentación de cada uno de los expertos. A partir de esto se evidencia que en el grupo el 100 % tiene un coeficiente de competencia alto lo que respalda la selección intencionada de los expertos por el autor.

Tabla 8. Datos obtenidos en el cálculo de K

Experto No.	AutoEval.	Kc	F1	F2	F3	F4	F5	F6	Ka	K	Grado
Experto 1	10	1,00	0,20	0,40	0,04	0,04	0,04	0,04	0,76	0,88	Alto
Experto 2	9	0,90	0,30	0,50	0,05	0,04	0,03	0,04	0,96	0,93	Alto
Experto 3	10	1,00	0,20	0,50	0,04	0,03	0,03	0,05	0,85	0,93	Alto
Experto 4	9	0,90	0,20	0,50	0,04	0,04	0,05	0,05	0,88	0,89	Alto
Experto 5	10	1,00	0,20	0,50	0,04	0,05	0,05	0,05	0,89	0,95	Alto
Experto 6	10	1,00	0,30	0,50	0,04	0,04	0,05	0,05	0,98	0,99	Alto
Experto 7	10	1,00	0,20	0,50	0,04	0,04	0,04	0,04	0,86	0,93	Alto
Experto 8	8	0,80	0,30	0,50	0,04	0,05	0,04	0,04	0,97	0,89	Alto
Experto 9	10	1,00	0,20	0,50	0,05	0,03	0,03	0,04	0,85	0,93	Alto
Experto 10	9	0,90	0,30	0,50	0,04	0,05	0,03	0,05	0,97	0,94	Alto
Experto 11	10	1,00	0,20	0,50	0,03	0,03	0,03	0,04	0,83	0,92	Alto
Experto 12	8	0,80	0,30	0,50	0,04	0,05	0,04	0,05	0,98	0,89	Alto
Experto 13	10	1,00	0,20	0,50	0,04	0,04	0,04	0,04	0,86	0,93	Alto
Experto 14	10	1,00	0,30	0,30	0,04	0,03	0,03	0,05	0,75	0,88	Alto
Experto 15	7	0,70	0,30	0,50	0,04	0,04	0,03	0,05	0,96	0,83	Alto
Experto 16	9	0,90	0,20	0,50	0,03	0,05	0,04	0,04	0,86	0,88	Alto
Experto 17	10	1,00	0,20	0,50	0,03	0,03	0,03	0,04	0,83	0,92	Alto
Experto 18	8	0,80	0,30	0,50	0,04	0,05	0,04	0,05	0,98	0,89	Alto

Además de este dato obtenido se puede decir que en el grupo el 44% son profesores de secundaria básica todos con una media de 15 años de experiencia en este nivel de enseñanza, los restantes 10 son especialistas universitarios con más de ocho años de experiencia en el trabajo y desarrollo de software educativos, dos de ellos posee el título de Doctor en Ciencias y otros nueve son Masters en Ciencias. Todos manifestaron su voluntariedad en la participación como expertos.

La aplicación de esta encuesta se realizó de manera individual y bajo total anonimato, con el objetivo de que ningún miembro del equipo fuera influenciado por la reputación de otro o por el peso que supone la opinión de la mayoría. Cada experto pudo defender sus argumentos tranquilamente sin el temor de que en caso de una equivocación, esta vaya a ser conocida por los otros. Esto posibilita obtener los verdaderos criterios de los participantes.

Escalamiento de Likert

Con el fin de ofrecer un resultado que visualice mejor las valoraciones de los expertos, se aplicó la escala de Likert, donde se otorga una puntuación entre 1 y 5 a cada ítem, como se muestra en la tabla 6.

Tabla 9. Asignación de valores usando Escalamiento de Likert [Elaboración propia]

Criterios de evaluación			Puntuación
Muy adecuado	Muy importante	Sí	5
Adecuado	Importante	Para la mayoría de los casos	4
Medianamente Adecuado	Medianamente importante	En Algunos Casos	3

Poco Adecuado	Poco Importante	Para la minoría de los casos	2
Inadecuado	Sin importancia	No	1

Luego se calculó un índice porcentual (IP) según la fórmula, que integra en un solo valor la aceptación del grupo de evaluadores sobre las características de la propuesta.

$$IP = \frac{5\ (\%\ de\ MA)+4(\%\ de\ A)+3\ (\%\ de\ MnA)+2(\%\ de\ PA)+1\ (\%\ de\ I)}{5} \qquad (4)$$

Donde:

- MA: Muy adecuado (o Muy importante)
- A: Adecuado (o Importante)
- MnA: Medianamente adecuado (o Medianamente importante)
- PA: Poco Adecuado (o Poco Importante)
- I: Inadecuado (Sin importancia)

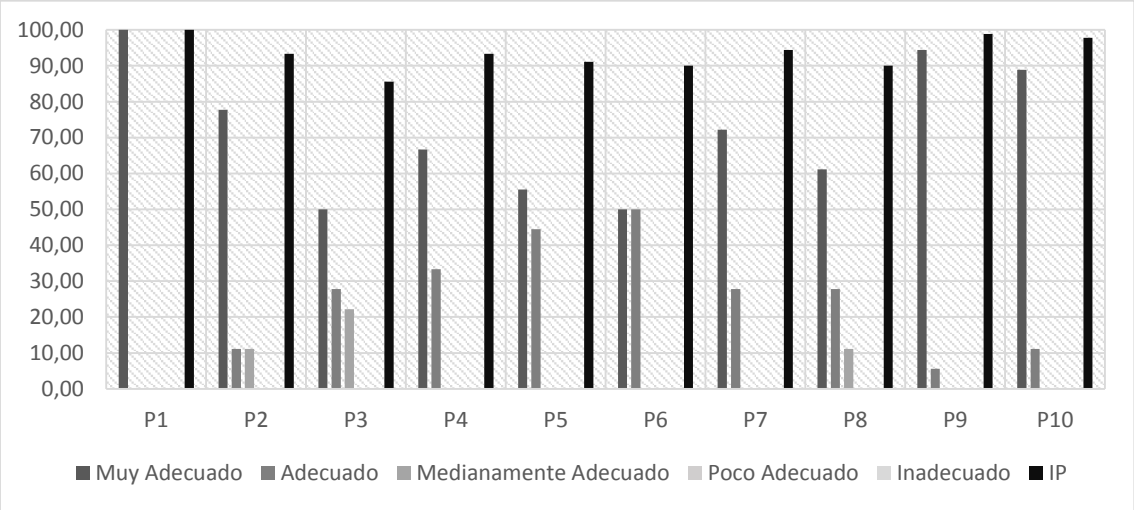


Figura 28. Datos obtenidos usando Escalamiento de Likert [Elaboración propia]

El procesamiento realizado a través del escalamiento de Likert, como se muestra en la **figura 28** evidencia que la aceptación de los expertos en la valoración de la propuesta sobrepasa el 90 % de IP a excepción de un solo ítems que tiene un IP = 85.56 %. Muestra la importancia y la aceptación de los reportes implementados para el módulo Resultados de la plataforma educativa Navigo, todos los reportes fueron valorados como importantes o muy importantes, al mismo tiempo que se valora de muy adecuado la necesidad de darle un seguimiento personalizado a los estudiantes. Por otra parte con valoraciones entre muy adecuado y adecuado están la necesidad de generar reportes utilizando minería de datos. A su vez se valora como muy adecuados la generación de los reportes que contribuyen al control de los resultados del aprendizaje y seguimiento personalizado de los estudiantes, ya sea de forma individual o por grupos.

3.4 Análisis del impacto económico y social de la solución

La solución propuesta incorpora a los hiperentornos gestionados por la plataforma educativa Navigo un proceso de extracción de conocimiento a partir de los registros almacenados por la plataforma educativa, lo cual será de gran aporte para los docentes que interactúen con estos productos durante el proceso de enseñanza y aprendizaje. Esta propuesta es aplicable a cualquier

producto educativo que tenga almacenados registros de la interacción de los usuarios y preferentemente sigan el modelo pedagógico y didáctico de hiperentorno de aprendizaje definido por los pedagogos cubanos, para esto solo habría que adaptar la fase de selección de los datos.

Las funcionalidades implementadas en el módulo Resultados no solo serán una herramienta importante para los docentes y alumnos, sino que también podrá poner en manos de los padres y familiares de los estudiantes el acceso a la información brindada por los reportes y tener un control del aprendizaje, además mantener un seguimiento del comportamiento de sus hijos en la interacción con el hiperentorno. Esto constituye un aporte a los proyectos sociales que se están llevando a cabo por el MINED con el objetivo de involucrar a la familia, la escuela y la comunidad en la educación de toda la sociedad cubana.

Desde el punto de vista económico el trabajo realizado aporta al cumplimiento de los Lineamientos de la Política Económica y Social del Partido y la Revolución, aprobados en el VI Congreso del Partido Comunista de Cuba, específicamente en la esfera Político Social, cumpliendo con el lineamiento 147 correspondiente a Educación plantea fortalecer el papel del profesor y lograr que los equipos y medios audiovisuales sean un complemento de la labor educativa del docente y garantizar el uso racional de los mismos. En este sentido la propuesta aporta a los docentes un mayor aprovechamiento de los datos almacenados extrayendo el conocimiento que puede ser de gran utilidad para el maestro en el proceso de enseñanza y aprendizaje una vez que se despliegue la plataforma educativa Navigo por todas las escuelas secundarias de Cuba.

La MD es un proceso costoso, ya que requiere de esfuerzos en la recolección de los datos, preparación de la información, la integración del software, la formulación del problema y la construcción del modelo de análisis. Todo este proceso es bastante caro, sin tener en cuenta los profesionales necesarios que según las encuesta publicadas en (KDnuggets, 2015), los salarios de un especialista en MD en América rondan los $58K[16] . Esto no es un problema para la institución que utilice la plataforma ya que no es necesario la intervención de un especialista en MD, pues a través de la solución propuesta un docente o cualquier otro tipo de usuario puede hacer uso de un proceso de KDD completo sin la necesidad del más mínimo conocimiento de MD, lo cual también influye en el ahorro de recursos. Con respecto a esto también se tiene que Navigo es un producto desarrollado totalmente en Cuba y con herramientas libres garantizando una soberanía tecnológica y se evita cualquier tipo de gastos por pago de licencias de software. Según KDnuggets el precio promedio de las licencias comerciales oscila por los 16 000 USD para herramientas como:

- IBM Analytical Decision Management
- IBM SPSS Modeler Professional
- IBM SPSS Statistics Professional
- SAS base
- SAS Enterprise Miner
- RapidMiner

Se evidencia el alto costo de las herramientas para el análisis y extracción de conocimiento en bases de datos, lo que dificulta su adquisición por parte de las entidades docentes cubanas, además

[16] Cincuenta y ocho mil USD al año.

de los considerables gastos que representan para la economía. A esto se debe sumar que la mayoría de las compañías son de origen estadounidense lo que constituye otra dificultad para su adquisición por motivos del bloqueo económico y financiero. Nada de esto constituye un impedimento para Navigo debido a que la herramienta que se utiliza para las tareas de minería es Weka bajó licencia GPL.

Conclusiones del capítulo

Luego de aplicar los métodos científicos con el objetivo de validar la propuesta de solución al problema planteado se concluye lo siguiente:

- Mediante la realización de las pruebas funcionales se pudo validar el cumplimiento de los requisitos de la solución informática implementada relacionados con el proceso de extracción de conocimiento a partir de la aplicación de MD y reportes estadísticos.
- La validación de los modelos propuestos demostró una gran aceptación por parte de los expertos, expresando los indicadores, interés, novedad, aplicabilidad, así como la comprensibilidad y la simplicidad con un índice porcentual por encima del 90%.
- La mayoría de los expertos coincidieron en que la solución contribuye al control de los resultados del aprendizaje y seguimiento de los estudiantes demostrándose el cumplimiento de la hipótesis planteada.
- La ejecución de tareas de MD es un proceso caro internacionalmente lo cual no es un problema para la solución implementada pues todo se realiza desde la plataforma educativa Navigo sin la necesidad de contratar especialistas en el tema, ni del pago de licencias de software gracias a la utilización de herramientas libres.

CONCLUSIONES

Los resultados obtenidos durante el desarrollo de la presente investigación permiten llegar a las siguientes conclusiones:

- La sistematización asociada al objeto de estudio demostró la importancia del uso de la MD en el descubrimiento de conocimiento, así como, el impacto significativo en los entornos educativos y la necesidad de que el módulo Resultados de la plataforma educativa Navigo incorpore nuevos reportes basados en la concepción didáctica diseñada por especialista del MINED y otros basados en técnicas de MD.
- La implementación de las funcionalidades, Trayectoria del estudiante, Análisis de contenidos, Historial del estudiante y Análisis constituyen un aporte considerable para la plataforma educativa Navigo que resuelve las deficiencias existentes en las versiones anteriores de las colecciones de hiperentornos existentes en Cuba.
- La solución informática propuesta incorpora a la plataforma educativa Navigo un conjunto de reportes basados en consultas y aplicación de técnicas de MD que contribuyen al control de los resultados del aprendizaje y seguimiento personalizado de cada estudiante o grupo en particular para darle cumplimiento a el objetivo trazado y solución a la problemática.
- Mediante la realización de las pruebas funcionales se pudo validar el cumplimiento de los requisitos de la solución informática implementada y según el criterio de la mayoría de los expertos, coincidieron en que la solución contribuye al control de los resultados del aprendizaje y seguimiento de los estudiantes y de esta manera queda demostrada la hipótesis de la investigación.
- El resultado obtenido es un aporte social para docentes, alumnos y familiares, además de contribuir al ahorro de la economía del país pues no es necesario la contratación de especialista en MD ni el pago de licencias de software para estas tareas, que resultan de alto valor internacionalmente.

RECOMENDACIONES

A partir de los resultados alcanzados, el estudio realizado y las conclusiones arribadas en la presente investigación el autor recomienda:

- Al grupo técnico y a la dirección del centro FORTES, analizar la incorporación de la solución propuesta en otros de los productos desarrollados en el centro.
- Continuar investigando sobre la MDE con el objetivo de aplicarla con otros enfoques como la recomendación y la adaptabilidad de los hiperentornos educativos.
- Continuar investigando en los temas de MD libre de parámetros para perfeccionar las técnicas utilizadas en la MDE y aplicarla a los productos educativos desarrollados en FORTES.

REFERENCIAS

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional space*: Springer.

Agraval, R., & Srikant, R. (1994). *'Fast Algorithms for Mining Association Rules in Large Data Bases.* Paper presented at the 20th Inter-national Conference on Very Large Databases, Santiagom.

Alvarez, S., Gonzalez, E., Pérez, Z., & Espinosa, I. (2007). Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando técnicas de minería de datos. http://eprints.rclis.org/10937/

Amaya Torrado, Y. K., Barrientos Avendaño, E., & Heredia Vizcaíno, D. J. (2014). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining, 1*(1), 3-17.

Balcázar, J. L. (2011). *Parameter-free Association Rule Mining with Yacaree.* Paper presented at the EGC.

Barujel, A. G. (2010). Diseño de entornos de aprendizaje. *Software Educativo, 24*.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*: John Wiley & Sons, Inc.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*: John Wiley & Sons.

Berthold, M., & Hand, D. J. (2003). *Intelligent data analysis: an introduction*: Springer Science & Business Media.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2008). *KNIME: The Konstanz information miner*: Springer.

Bodon, F. (2010). *A fast apriori implementation.* Paper presented at the Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI'03).

Brito, R. (2008). *Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico" José Antonio Echeverría.* (Tesis de maestría), ISPJAE, La Habana, Cuba.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*: Prentice-Hall, Inc.

CALIDAD-UCI. (2014). Informe de Cierre del Pruebas de Liberación. La Habana, Cuba: UCI.

Casañola, Y. T. (2014). Acta de liberación del laboratorio de Pruebas-UCI. La Habana: UCI.

Crows, T. (1999). Introduction to data mining and knowledge discovery. *Two Crows Corporation, 36*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Del Toro, M. R. (2006). *Modelo de diseño didáctico de hiperentornos de enseñanza–aprendizaje desde una concepción desarrolladora.* (Doctoral dissertation), Instituto Superior Pedagógico "Enrique J. Varona" La Habana, Cuba.

Dios, M. L. (2015). *Investigación y recogida de información de mercados: Identificación de variables de estudio y desarrollo del trabajo de campo*: Ideaspropias Editorial SL.

Fayyad, U., Haussler, D., & Stolorz, P. E. (1996). *KDD for Science Data Analysis: Issues and Examples.* Paper presented at the KDD.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. *MIT Press*.

Fayyad, U. M., Wierse, A., & Grinstein, G. G. (2002). *Information visualization in data mining and knowledge discovery*: Morgan Kaufmann.

Félix, L. C. M. (2002). Data mining: torturando a los datos hasta que confiesen. *Coordinador del programa de Data mining*.

Ferri, C., Flach, P., & Hernández-Orallo, J. (2002). *Learning decision trees using the area under the ROC curve.* Paper presented at the ICML.

Galindo, Á. J., & García, H. Á. (2010). Minería de Datos en la Educación. Universidad Carlos III de Madrid: Madrid, España.

García-Saiz, D. (2011). *Data Mining applied to educational environments: research and implementation of parameter-free techniques.* (Tesis de Maestría).

Garcıa-Saiz, D., & Zorrilla, M. (2011). Hacia la minería de datos sin parámetros y su aplicación en el campo educativo.

García, L., & Fernández, S. J. (2008). Procedimiento de aplicación del trabajo creativo en grupo de expertos. *Ingeniería Energética, 29*(2), 46-50.

García, P. G. (2008). *Aprendizaje evolutivo de reglas difusas para descripción de subgrupos*: Universidad de Granada.

Garre, M., Cuadrado, J. J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Revista Española de Innovación, Calidad e Ingeniería del Software, 3*(1), 6-22.

Gelles, I. B., del Toro, M., Valle, P. R., & Armenteros, I. R. (2011). Educación y tecnologías de la información y las comunicaciones: una mirada desde la formacion del docente. *Educación Cubana*.

Hall, M., Frank, E., Bouckaert, R. R., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). WEKA Manual for Version 3-7-8: January.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (Third Edition ed.): The Morgan Kaufmann Series in Data Management Systems.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*: MIT press.

Hartigan, J. A. (1975). Clustering algorithms. *John Wiley & Sons, New York*.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*, 100-108.

Hasim, N., & Haris, N. A. (2015). *A study of open-source data mining tools for forecasting.* Paper presented at the Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication.

Hasperué, W. (2012). *Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas.* (Tesis Doctoral en Ciencias Informáticas), UNIVERSIDAD NACIONAL DE LA PLATA FACULTAD DE INFORMÁTICA, Buenos Aires, Argentina.

Hernández-Orallo, J., Ramírez, M. J. Q., & Ferri, C. R. (2004). Introducción a la Minería de Datos. *Editorial Pearson Educación SA, Madrid*.

Herrero, J. G., & López, J. M. M. (2006). *Técnicas de análisis de datos: Aplicaciones prácticas usando Microsoft Excel y Weka*. Universidad Carlos III Madrid.

Hilbert, M. R., & Peres, W. (2009). *La Sociedad de la Información en América Latina y el Caribe: Desarrollo de las Tecnologías y Tecnologías para el Desarrollo* (Vol. 98): United Nations Publications.

Hu, H.-W., Chen, Y.-L., & Tang, K. (2009). A dynamic discretization approach for constructing decision trees with a continuous label. *Knowledge and Data Engineering, IEEE Transactions on, 21*(11), 1505-1514.

Hua, H., & Zhao, H. (2009). *A Discretization Algorithm of Continuous Attributes Based on Supervised Clustering.* Paper presented at the Pattern Recognition, 2009. CCPR 2009. Chinese Conference on.

Hunyadi, D. (2011). *Performance comparison of Apriori and FP-Growth algorithms in generating association rules.* Paper presented at the Proceedings of the European Computing Conference.

ISO9000. (2001).

Jiménez, M. G., & Álvarez, A. (2010). Análisis de datos en WEKA–pruebas de selectividad.

Johnson, L., Becker, S., Estrada, V., & Freeman, A. (2015). NMC Horizon Report: 2015 Higher Education Edition. *Austin, Texas:The New Media Consortium*.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms 2nd* (P. N. J. Wiley-Interscience Ed.). Estados Unidos: John Wiley & Sons.

KDnuggets. (2014a). What Analytics, Data Mining, Data Science software/tools you used in the past 12 months for a real project Poll. Retrieved 20 de enero, 2015, from http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html

KDnuggets. (2014b). What main methodology are you using for your analytics, data mining, or data science projects? Retrieved 10 de mayo, 2015, from http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

KDnuggets. (2015). Annual Income/Salary for Analytics, Data Mining, Data Science Professionals Poll. Retrieved 18 mayo, 2015, from http://www.kdnuggets.com/polls/2015/salary-analytics-data-science-data-mining.html

Knime. (2015). Open for Innovation. from http://www.knime.org

Lara, A. T. (2011). *Marco de descubrimiento de conocimiento para datos estructuralmente complejos con énfasis en el análisis de eventos en series temporales.* (Doctoral dissertation), Universidad Politécnica de Madrid, España.

Marqués, P. (1996). El software educativo. *Comunicación educativa y Nuevas Tecnologías*, 119-144.

Marqués, P., Martinez, A. M. G., López, K. A., Peralta, E. J., & Zuñiga, S. F. (2009). El software educativo: Universidad Autónoma de Barcelona.

Martín, C. E. R., & Arias, Y. E. P. (2015). *Identificación de sismos similares haciendo uso de técnicas de Minería de Datos.* (Tesina Diplomado de Gestión de Información con Técnicas de Minería de Datos), Instituto de Cibernética Matemática y Física ICIMAF, La Habana, Cuba.

Microsoft. (2015). Conceptos de minería de datos. Retrieved 7 de abril, 2015, from https://msdn.microsoft.com/es-es/library/ms174949.aspx

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(5), 431-443.

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo.* (Maestría en Ingeniería de Software), Universidad Nacional de la Plata, Argentina.

Moine, J. M., Haedo, A., & Gordillo, S. (2011). *Estudio comparativo de metodologías para Minería de Datos.* Paper presented at the XIII Workshop de Investigadores en Ciencias de la Computación.

Mor, E., & Minguillón, J. (2004). *E-learning personalization based on itineraries and long-term navigational behavior*. Paper presented at the Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters.

Morales, C. R., Soto, S. V., & Martínez, C. H. (2005). Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005*, 49-56.

Orallo, J. H., Quintana, M. J. R., & Ramírez, C. F. (2006). *Práctica de Minería de Datos Introducción al Weka*. Paper presented at the Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software, Universitat Politècnica de València.

Park, J. S., Chen, M.-S., & Yu, P. S. (1995). *An effective hash-based algorithm for mining association rules* (Vol. 24): ACM.

Peng, Y., Kou, G., Shi, Y., & Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making*, *7*(04), 639-682.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462.

Peña, A., Domínguez, R., & Medel, J. d. J. (2009). Educational data mining: a sample of review and study case. *World Journal On Educational Technology*, *1*(2), 118-139.

Pérez Araujo, L. I. (2013). *Metodología para la gestión de contenidos en el desarrollo de hiperentornos de aprendizaje diseñados por el Ministerio de Educación de Cuba.* (Maestría en Gestión de Proyectos), Universidad de las Ciencias Informática, La Habana, Cuba.

Pérez, I., & León, B. (2007). Lógica difusa para principiantes. *Publ. UCAB. Caracas*.

Piatetsky, G., & Frawley, W. (1991). *Knowledge discovery in databases*: MIT press.

Ponce, R. V., & Alcaraz, J. L. G. (2013). Evaluación de Tecnología utilizando TOPSIS en Presencia de Multi-colinealidad en Atributos:¿ Por qué usar distancia de Mahalanobis? *Revista Facultad de Ingeniería Universidad de Antioquia*(67), 31-42.

Prekopcsak, Z., Makrai, G., Henk, T., & Gaspar-Papanek, C. (2011). *Radoop: Analyzing big data with rapidminer and hadoop.* Paper presented at the Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011).

Pyle, D. (1999). *Data preparation for data mining* (Vol. 1): Morgan Kaufmann.

R-Project. (2015). The R Foundation. Retrieved 10 de enero, 2015, from http://www.r-project.org/

RapidMiner. (2015). RapidMiner. from https://rapidminer.com

Report, I. R. a. D. U. (2010). 2010 Report: Open Source Data Mining Software Evaluation. USA.

Rizzo, C. L. (2009). El software educativo en el contexto de la escuela cubana. *La Habana: Pueblo y educación*.

Rodríguez, C. O. C., Blanco, D. M., Rodríguez, Y. P., Verdecia, R. R., Ricardo, G. C., Oliva, Y. T., . . . Mulet, A. R. (2011). EL DESARROLLO DE SOFTWARE EDUCATIVO SIN COSTO DE PROGRAMACIÓN.¿ UTOPÍA O REALIDAD? In E. C. M. d. Educación (Ed.). La Habana.

Rodríguez, E. P. (2014). *Descubrimiento de conocimiento a partir de la relación rasgos de la personalidad-rendimiento laboral en proyectos informáticos*. (Maestría en Gestión de Proyectos), Universidad de las Ciencias Informáticas, La Habana, Cuba.

Rodríguez, L. A. R., Pons, E. S., López, A., Mora, N. M., Hernández, R., Morales, Y. C., & Suárez, P. A. (2005). *Módulo Resultados de la Colección Futuro de Preuniversitario*. Paper presented at the Pedagogía 2005, Cuba.

Rodríguez, L. R. (2010). *Concepción Didáctica del Software Educativo como instrumento mediador para un aprendizaje desarrollador*. (Doctoral dissertation), Universidad de Ciencias Pedagógicas "Félix Varela y Morales", Santa Clara, Cuba.

Román, A. B., Sánchez-Guzmán, D., & García, R. (2014). Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. *Lat. Am. J. Phys. Educ. Vol, 7*(4), 662.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications, 33*(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40*(6), 601-618.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). *Handbook of educational data mining*: CRC Press.

Saldaña, J. F. R., & Flores, R. G. (2005). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías, 8*(26), 37.

Sánchez, D. R. R. (2005). *Selección de Atributos mediante proyecciones*. (Doctoral dissertation), Universidad de Sevilla, España.

Sarasa, R. B., Suárez, A. R., & Sánchez, R. A. (2008). *Desarrollo de un proceso de KDD en el ámbito docente: Preparación de los datos*. Paper presented at the 14 Convención científica de ingeniería y arquitectura, Instituto Superior Politécnico José Antonio Echeverría CUJAE. La Habana. Cuba.

Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases.

Soto, C., & Jiménes, C. (2011). APRENDIZAJE SUPERVISADO PARA LA DISCRIMINACIÓN Y CLASIFICACIÓN DIFUSA SUPERVISED LEARNING FOR FUZZY DISCRIMINATION AND CLASSIFICATION. *Dyna, 169*, 27.

Suárez, A. R., Sarasa, R. B., & Sánchez, R. A. (2008). *Resultados de la aplicación de algoritmos de Minería de Datos a la Gestión Docente*. Paper presented at the 14 Convención científica de ingeniería y arquitectura, Instituto Superior Politécnico José Antonio Echeverría CUJAE. La Habana. Cuba.

Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications* (Vol. 29): Springer Science & Business Media.

Tanna, P., & Ghodasara, Y. (2014). Using Apriori with WEKA for Frequent Pattern Mining. *International Journal of Engineering Trends and Technology (IJETT), 12*(3).

Thuraisingham, B. (1998). *Data mining: technologies, techniques, tools, and trends*: CRC press.

Weka. (2015). Weka 3: Data Mining Software in Java. from http://www.cs.waikato.ac.nz/ml/weka/

Winters, T. d. (2006). *Educational data mining: collection and analysis of score matrices for outcomes-based assessment.* (Doctoral dissertation), University of California, Riverside.

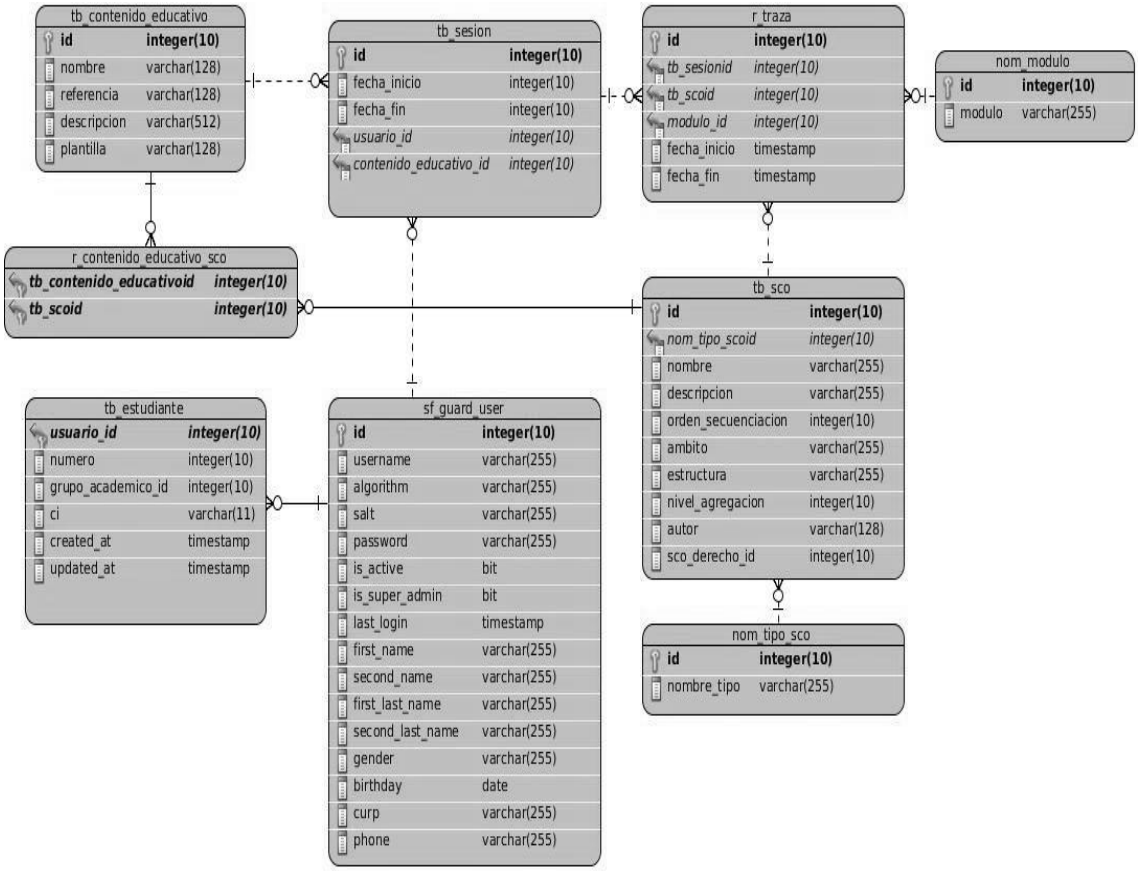
Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. Paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Witten, I., & Frank, E. (2005). The Morgan Kaufmann series in data management systems: San Francisco, CA: Morgan Kaufmann Publishers.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (Second Edition ed.): Morgan Kaufmann.

Zorrilla, M., García-Saiz, D., & Balcázar, J. (2011). Towards parameter-free data mining: Mining educational data with yacaree.

ANEXO 1: Fragmento del modelo de Base de Datos de la plataforma educativa Navigo



ANEXO 2: Ejemplar de la encuesta para evaluar los modelos descriptivos obtenidos

CONSULTA A EXPERTOS

Estimado compañero(a): La presente encuesta forma parte de una investigación dirigida al análisis de los registros generados en Hiperentornos educativos. Como Experto conocedor de los temas **Software Educativo** y/o **Aprendizaje Escolar** solicitamos que ofrezca su más sincera y precisa valoración sobre aspectos que se pondrán a su consideración sobre la **plataforma educativa Navigo**, con lo cual contribuirá a evaluar los modelos obtenidos después de aplicar minería de datos a los registros de la plataforma.

Nota: El objetivo de esta encuesta es solamente investigativa. El responsable de esta encuesta se compromete a mantener total privacidad de la información recopilada.

1. Datos generales del encuestado:

Institución y Dpto. donde labora: _____
Título universitario: _____
Categoría científica: _____
Categoría docente:_____
Años de experiencia en la docencia: _____
Frecuencia con que utiliza el software educativo ( ) alta ( ) media ( ) baja

2. Evaluación de los modelos

Evalúe los siguientes elementos según el nivel de cumplimiento que usted le confiera a los reportes generados por la plataforma educativa Navigo a partir de la aplicación de técnicas de Minería de Datos como Agrupamiento y Reglas de asociación (en el archivo adjunto se describen cada uno de los reportes generados), donde:

(1) significa poco valor.

(5) significa el máximo valor posible.

Criterios	1	2	3	4	5
Interés: intenta medir la capacidad del modelo de suscitar la atención del usuario.					
Novedad: relacionado con la capacidad de un modelo de sorprender al usuario con respecto al conocimiento previo que tenía sobre determinado problema.					
Comprensibilidad o Inteligibilidad:					
Simplicidad: se basa en el tamaño o complejidad del modelo.					
Aplicabilidad: capacidad de ser utilizado con éxito en el contexto real donde va ser aplicado.					

3. Si desea exponer cualquier otra opinión, por favor, expréselo en el espacio disponible a continuación.

MUCHAS GRACIAS POR SU COLABORACIÓN

ANEXO 3: Ejemplar de la encuesta para obtener los criterios de los expertos sobre la solución propuesta

CONSULTA A EXPERTOS

Estimado compañero(a): La presente encuesta forma parte de una investigación dirigida al análisis de los registros generados en Hiperentornos educativos. Como Experto conocedor de los temas **Software Educativo** y/o **Aprendizaje Escolar** solicitamos que ofrezca sus más sinceras y precisas valoraciones sobre aspectos que se pondrán a su consideración sobre la **plataforma educativa Navigo**, con lo cual contribuirá a evaluar si las reportes implementados en el módulo Resultados contribuyen al control y seguimientos del aprendizaje de los estudiantes por parte de los docentes después de extraer y analizar los conocimientos implícitos en los registros de la interacción de los estudiantes con la plataforma.

Nota: El objetivo de esta encuesta es solamente investigativa. El responsable de esta encuesta se compromete a mantener total privacidad de la información recopilada.

Datos generales del encuestado:

Institución y Dpto. donde labora: _____
Título universitario: _____
Categoría científica: _____
Categoría docente: _____
Años de experiencia en la docencia: _____
Años de experiencia en el trabajo con software educativo: _____

PREGUNTAS

1. **¿Considera usted importante el módulo Resultados de la plataforma educativa Navigo?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

2. **Valorar si el seguimiento personalizado del comportamiento de cada estudiante o grupo de estudiantes es necesario para la correcta evaluación de los mismos.**

___Muy adecuados ___Adecuados ___Medianamente Adecuados ___Poco adecuados ___Inadecuados

3. **¿Qué importancia le atribuye usted a la generación del reporte Trayectoria del estudiante para el control y seguimiento del aprendizaje de los estudiantes en la plataforma educativa Navigo?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

4. **¿Qué importancia le atribuye usted a la generación del reporte Análisis de contenidos para el control y seguimiento del aprendizaje de los estudiantes en la plataforma educativa Navigo?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

5. **¿Qué importancia le atribuye usted a la generación del reporte Historial del estudiante para el control y seguimiento del aprendizaje de los estudiantes en la plataforma educativa Navigo?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

6. **¿Qué importancia le atribuye usted a la generación del reporte Análisis integral para el control y seguimiento del aprendizaje de los estudiantes en la plataforma educativa Navigo?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

7. **¿Qué importancia le atribuye usted a la generación de los reportes del módulo Resultados de la plataforma educativa Navigo utilizando técnicas de agrupamiento?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

8. **¿Qué importancia le atribuye usted a la generación de los reportes del módulo Resultados de la plataforma educativa Navigo utilizando técnicas de reglas de asociación?**

___Muy importante ___Importante ___ Medianamente importante ___Poco Importante

___Sin importancia

9. **Valorar si es posible que los reportes descriptivos del módulo Resultados de la plataforma educativa Navigo contribuyan al control de los resultados del aprendizaje.**

___Sí ___ Para la mayoría de los casos ___En algunos casos

___ Para la minoría de los casos ___ No

10. **Valorar si es posible que los reportes descriptivos del módulo Resultados la plataforma educativa Navigo contribuyan seguimiento personalizado de cada estudiante o grupo en particular por parte de los docentes en la plataforma educativa Navigo.**

___Sí ___ Para la mayoría de los casos ___En algunos casos

___ Para la minoría de los casos ___ No

Como parte del método de procesamiento de los datos obtenidos por medio de la presente encuesta, se necesita caracterizar estadísticamente la competencia del conjunto de expertos del cual usted forma parte, por lo que finalmente se le pide ayude respondiendo lo más fielmente posible al siguiente **TEST DE AUTOVALORACIÓN DEL CONSULTADO**:

a) Evalúe su nivel de dominio acerca de la esfera sobre la cual se le consultó marcando con una cruz sobre la siguiente escala (1: dominio mínimo; 10: dominio máximo)

1	2	3	4	5	6	7	8	9	10

b) Evalúe la influencia de las siguientes fuentes de argumentación en los criterios valorativos aportados por usted.

Fuentes de argumentación	Grado de influencia de las fuentes de argumentación.		
	Alto	Medio	Bajo
Análisis teóricos realizados por usted			
Su propia experiencia			
Trabajos de autores nacionales			
Trabajos de autores extranjeros			
Su conocimiento del estado del problema en el extranjero			
Su intuición			

MUCHAS GRACIAS POR SU COLABORACIÓN.

ANEXO 4: Valoraciones de los expertos sobre los modelos obtenidos

Expertos	Interés	Novedad	Comprensibilidad	Simplicidad	Aplicabilidad
E1	5	5	4	5	5
E2	5	5	4	4	5
E3	5	4	5	5	5
E4	5	5	5	4	5
E5	5	5	5	5	5
E6	4	5	5	5	5
E7	5	5	4	5	4
E8	4	5	5	5	5
E9	5	4	5	5	5
E10	5	5	4	5	5
E11	5	5	5	4	5
E12	4	5	5	4	5
E13	5	4	4	5	5
E14	3	4	5	5	3
E15	5	5	5	5	5
E16	5	4	5	5	5
E17	5	5	4	4	5
E18	5	4	5	5	5
E19	4	5	5	5	5
E20	5	5	5	5	4
E21	4	5	5	4	4
E22	5	5	4	4	5
E23	3	5	4	4	3
Suma	**106**	**109**	**107**	**107**	**108**
Media	**4,61**	**4,74**	**4,65**	**4,65**	**4,70**
Desviación	**0,64**	**0,44**	**0,48**	**0,48**	**0,62**
IP	**92,17**	**94,78**	**93,04**	**93,04**	**94,41**

ANEXO 5: Valoraciones de los expertos sobre la propuesta

Preguntas	EXPERTOS																	
	E 1	E 2	E 3	E 4	E 5	E 6	E 7	E 8	E 9	E 10	E 11	E 12	E 13	E 14	E 15	E 16	E 17	E 18
1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5
2	5	4	4	5	5	5	5	3	5	5	5	5	3	5	5	5	4	5
3	4	5	3	5	3	4	3	5	5	4	5	5	4	5	5	5	3	4
4	5	5	4	5	5	5	4	5	5	5	4	4	4	4	5	5	5	5
5	5	5	4	5	4	5	4	4	4	5	5	5	5	5	5	5	4	5
6	5	5	4	5	5	4	5	4	5	5	5	5	5	4	5	5	4	5
7	5	5	5	4	4	5	5	5	5	5	5	4	5	5	5	4	5	5
8	5	5	5	4	5	3	5	5	5	3	5	4	5	5	5	4	4	5
9	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	4	5
10	5	5	4	5	5	5	5	5	5	5	5	4	5	5	5	5	4	5