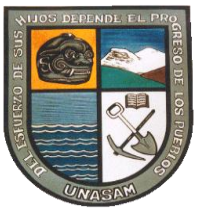


UNIVERSIDAD NACIONAL "SANTIAGO ANTUNEZ DE MAYOLO"

ESCUELA DE POSTGRADO

UTILIZACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE RASGOS DE COMPORTAMIENTO EN PROCESOS DE APRENDIZAJE COLABORATIVO EN MODELOS DE E-LEARNING Y B-LEARNING

Tesis para optar el grado de Doctor
en Ciencias e Ingeniería de la Computación

PILAR SARA NUÑEZ BLAS

Asesor: **Dr. ESMELIN NIQUIN ALAYO**

Huaraz – Perú

2016

Nº. Registro: TE0021

MIEMBROS DEL JURADO

Doctor Alexander Pacheco Castillo — Presidente

_____

Doctor Edwin Johny Asnate Salazar — Secretario

_____

Doctor Esmelin Niquin Alayo — Vocal

_____

ASESOR

Doctor Esmelin Niquin Alayo

Agradecimiento

A mi asesor Dr. Esmelin Niquin Alayo a quien debo agradecer la propuesta del tema, por su paciencia, apoyo, dirección y estímulo en todo momento. Así como al Dr. Ángel Cobo Ortega por el aporte y sus sabios consejos en la ejecución de esta investigación.

A los colegas que me brindaron su apoyo desinteresado en todo momento de mis estudios y en el desarrollo de este trabajo

El presente trabajo está dedicado a:

A mi esposo Franklin, por el apoyo y aliento constante que juntos a mis hijos Ingrid y Franklin me supieron brindar; ellos son mi fuente de inspiración en mi deseo de seguir superándome.

Pilar

INDICE

		Página
RESUMEN		ix
ABSTRACT		x
I.	INTRODUCCIÓN	11
	1.1. OBJETIVOS	12
	1.2. HIPÓTESIS	13
	1.3. VARIABLES	14
II.	MARCO TEÓRICO	15
	2.1. ANTECEDENTES	15
	2.2. BASES TEÓRICAS	18
	2.2.1. Las Tic en la educación	18
	2.2.2. E-learning y B-learning	20
	2.2.3. Sistema E-learning y B-learning en el Perú	23
	2.2.4. Rasgos de aprendizaje en los modelos E-learning y B-learning	24
	2.2.5. Gestión de cursos virtuales	25
	2.2.6. Minería de datos	27
	2.2.7. Minería de datos educativos (EDM)	34
	2.2.8. Proceso de análisis jerárquico (AHP)	43
	2.2.9. Herramienta weka	45
	2.2.10. Herramienta expert choice	47

2.3. DEFINICIÓN DE TÉRMINOS 48

III. METODOLOGIA 52

3.1. TIPO Y DISEÑO DE INVESTIGACIÓN 53

3.1.1. De acuerdo a la orientación 53

3.1.2. De acuerdo a la técnica de contrastación 53

3.2. PLAN DE RECOLECCIÓN DE LA INFORMACIÓN Y/O DISEÑO ESTADÍSTICO 53

3.2.1 Población 53

3.2.2. Muestra 53

3.3. INSTRUMENTO(S) DE RECOLECCIÓN DE LA INFORMACIÓN 55

3.4. PLAN DE PROCESAMIENTO Y ANÁLISIS ESTADÍSTICO DE LA INFORMACIÓN 56

IV. RESULTADOS 66

4.1. MEDICIÓN DE NIVELES DE INTERACTIVIDAD UTILIZANDO EL PROCESO DE ANÁLISIS JERÁRQUICO (AHP) 66

4.1.1. Modelación del problema 68

4.1.2. Evaluación y priorización de alternativas y criterios 70

4.1.3. Síntesis del problema 73

4.1.4. Análisis de sensibilidad 78

4.2. APLICACIÓN DE LAS TÉCNICAS DE MINERÍA DE DATOS SOBRE LOS INDICADORES MODELO B-LEARNING 79

4.3. APLICACIÓN DE LAS TÉCNICAS DE MINERÍA DE DATOS SOBRE LOS INDICADORES MODELO E-LEARNING 113

4.4. COMPARACIÓN DE LOS MODELOS B-LEARNING Y L E-LEARNING 132

V. DISCUSIÓN 137

VI. CONCLUSIONES 140

VII. RECOMENDACIONES 142

VIII. REFERENCIAS BIBLIOGRAFICAS 144

RESUMEN

El propósito fundamental en la presente investigación es usar técnicas de Minería de Datos y la metodología de Proceso de Análisis Jerárquico (AHP) para identificar los rasgos de comportamientos en el aprendizaje colaborativo, así como la interactividad, el seguimiento de las actividades, la participación en los foros, a partir de información extraída de los logs que quedan en los servidores de la plataforma Moodle, en las modalidades E-learning y B-learning.

Se presenta detalladamente el procesamiento y análisis de los datos aplicando técnicas de Clustering con los algoritmos SimpleK-Means y (Expectation Maximization) EM, los pesos generados de todos los criterios y alternativas de la decisión multicriterio, identificando un índice cuantitativo de la interactividad del estudiante que varía en términos de las relevancia de cada profesor, con el apoyo de Expert Choice, Weka, softwares de AHP, Minería de Datos.

Finalmente, se presenta los resultados que nos ha permitido afirmar que los patrones de comportamiento del estudiante en los modelos E-learning y B-learning son muy disparejos. Por lo tanto, resulta muy útil y de interés para los profesores, del mismo modo tener información objetiva sobre del uso de sus cursos.

Palabras Claves: Moodle, E-learning y B-learning, aprendizaje colaborativo, Minería de Datos, (AHP)

ABSTRACT

The present purpose research is to use techniques Data Mining and Methodology Process Hierarchical Analysis (AHP) the key paragraph in the identify behavioral traits in collaborative learning and interactivity, monitoring activities, Participation in forums, information extracted from records remaining servers in the Moodle platform, the modalities of E-learning and B-learning.

It presents in detail the Processing and Data Analysis using Clustering Techniques with algorithms SimpleK-half and (Expectation Maximization) MS, generated weights of all criteria and Alternatives multicriteria decision, identifying the United Nations quantitative index of interactivity student varies in terms of the relevance of each teacher, with the support of expert Choice, Weka, AHP software, data Mining.

Finally, the results presented to us has allowed Stating that patterns of student behavior on E-learning Models to d-learning child Very uneven. Therefore, very useful results Search and interest to teachers, just as objective to have information about the use of their courses.

Keywords: Moodle, E-learning and B-learning, Collaborative learning, data mining, (AHP)

I INTRODUCCIÓN

La aplicación de técnicas de descubrimiento de conocimiento y minería de datos en sistemas educacionales basados en web es un área de investigación muy activa y con indudables aplicaciones prácticas. En el proceso de enseñanza-aprendizaje desempeñan un papel importante las tecnologías de la información, y en particular los entornos virtuales de enseñanza y las herramientas de la Web social. Diversos estudios concluyen que una mezcla en el aprendizaje, *face-to-face* y el aprendizaje E-learning resulta muy adecuada para implementar nuevos modelos de enseñanza. Este tipo de enseñanza, resulta especialmente efectiva, por lo que se tiende a defender modelos flexibles, semipresenciales o de enseñanza mixta que se agrupan bajo el término de *blended learning*, (Nowell, 2011).

Las técnicas de minería de datos, concebidas como metodologías para el análisis inteligente de los datos, ayudan a descubrir conocimiento a partir de la información de utilización de los sistemas de *E-learning y B-learning*. Permiten descubrir relaciones entre actividades y estudiantes, ayudan a obtener una mayor retroalimentación de la enseñanza, conocer cómo los

estudiantes aprenden, evaluar a los estudiantes por sus patrones de navegación, clasificar los estudiantes en grupos que responden a comportamientos similares, etc. (Romero et al, 2005).

1.1 OBJETIVOS

1.1.1. Objetivo General

Usar las técnicas de Minería de Datos y el Proceso de Análisis Jerárquico para identificar los rasgos de comportamiento, así como para medir el grado de interactividad de los participantes en procesos de aprendizaje colaborativo apoyados sobre los modelos E-learning y B-learning.

1.1.2. Objetivos Específicos

- Automatizar el seguimiento de las actividades desarrolladas por los participantes en los modelos E-learning y B-learning; en particular la participación en los foros.
- Usar técnicas de la minería de datos para la extracción de conocimiento en los modelos E-learning y B-learning.
- Clasificar o agrupar alumnos en base a sus características comportamentales.
- Realizar una medición del nivel de interactividad de los estudiantes, así como analizar la frecuencia y periodicidad

de accesos a las actividades y foros mediante un proceso analítico-jerárquico.

1.2. HIPÓTESIS

Usando técnicas apropiadas de minería de datos y el proceso de Análisis Jerárquico sobre los datos extraídos de los modelos E-learning y B-learning, será posible la identificación de rasgos de comportamiento de los usuarios de esos modelos; así como medir el grado de interactividad en el aprendizaje colaborativo.

1.3. VARIABLES

1.3.1. Variables Independientes

Informaciones tomadas de las bases de datos de los modelos E-learning y B-learning, obtenidos directamente a través de las opciones de administración de la plataforma Moodle.

1.3.2. Variables Dependientes

Rasgos de comportamiento de los estudiantes en los modelos E-learning y B-learning, e indicadores de interactividad global de los estudiantes.

II. MARCO TEÓRICO

2.1. ANTECEDENTES

En los últimos años, los investigadores han venido trabajando en diversos métodos para ayudar a los instructores y administradores a mejorar los sistemas de aprendizaje En los modelos de E-learning y B-learning se acumula una gran cantidad de información que es muy valiosa para analizar el comportamiento de los estudiantes y podría crear una "mina de datos" educativos. En estos sistemas se puede registrar cualquier actividad de los estudiantes involucrados, la producción de materiales de contenido, preparar las tareas y exámenes, participar en discusiones, y permiten el aprendizaje colaborativo con foros, chats, áreas de almacenamiento de archivos, servicios de noticias, etc. Por lo general también proporcionan una base de datos que almacena diferentes tipos de información: datos personales sobre los usuarios, sus resultados académicos, accesos a las diferentes actividades, distribuciones temporales de los accesos. Sin embargo, debido a la gran cantidad de datos que los sistemas E-learning y B-learning generan todos los días, resultan muy

difíciles de manejarlos manualmente, por lo que el uso de herramientas informáticas de gestión y análisis automático resulta imprescindible si realmente se desea generar nuevo conocimiento a partir de esos datos. Dentro de esta masa de datos hay información oculta de gran importancia que se podría llegar a descubrir con técnicas de Minería de datos.

Cabe destacar que la Minería de Datos (MD) es una etapa dentro de un proceso más amplio llamado Descubrimiento de Conocimiento en Bases de Datos (KDD, Knowledge Discovery in Databases), que puede verse como la extracción automática de patrones implícitos e interesantes desde grandes colecciones de datos. La minería de datos es un área multidisciplinar en la que convergen varios paradigmas de computación: la construcción de árboles de decisión, la inducción de regla general, las redes neuronales artificiales, el aprendizaje bayesiano, programación lógica, los algoritmos estadísticos, etc.

En una reciente publicación (Liao, Chu, & Hsiao, 2012) se presentan los resultados de una revisión de artículos académicos y científicos publicados en las últimas décadas relacionados con la aplicación de las diferentes técnicas de la minería de datos, y cómo esas técnicas y aplicaciones se han desarrollado en el periodo mencionado con respecto a tres áreas diferentes: tipo de conocimiento, tipo de análisis y tipo de arquitectura

Algunas de las tareas de minería de datos y métodos más útiles son la visualización, el agrupamiento, clasificación y las reglas de asociación (Brusilovsky 2008**).**

En los modelos E-learning y B-learning las técnicas de minería de datos se han aplicado para la evaluación de los resultados del aprendizaje de los estudiantes y para una gran variedad de aplicaciones prácticas complementarias. Las principales aplicaciones de las técnicas de minería de datos en educación, (García, Romero, Ventura, 2011), describe una herramienta de minería de datos que utiliza la asociación descartar la minería y el filtrado colaborativo con el fin de hacer recomendaciones a los instructores acerca de cómo mejorar el aprendizaje electrónico cursos. (Srivastava 2000), sistemas recomendadores, de clasificación de alumnos y de contenidos, construcción adaptativa de planes de enseñanza, descubrimiento de relaciones entre actividades, diagnóstico incremental de los estudiantes. Finalmente, en la actualidad se está trabajando sobre la aplicación de minería de datos en educación basada en web, teniendo en cuenta:

- Integración de algoritmos de minería de datos dentro de las propias herramientas de construcción y mantenimiento de los cursos, para la mejora automática de los sistemas.
- Integración de algoritmos de minería en los interfaces de usuario de los cursos y de las plataformas dentro de sistemas agentes recomendadores, para realización de sugerencias sobre rutas, actividades, etc.
- Desarrollo de algoritmos de minería de datos específicos para problemas relacionados con la enseñanza y el aprendizaje a distancia utilizando entornos hipermedia adaptativos y sistemas tutores inteligentes basados en web.

- Desarrollo de cuadros de mando (dashboards) con información e indicadores clave y dirigidos a los diferentes usuarios de los sistemas: estudiantes, profesores y administradores.

2.2. BASES TEÓRICA

2.2.1. Las TIC en la Educación

Las nuevas Tecnologías de la Información y de la Comunicación (TIC) han evolucionado espectacularmente en los últimos años, han llegado a ser uno de los pilares básicos de la sociedad, debido especialmente a su capacidad de interconexión a través de la Red.

El sistema educativo no puede quedar al margen de los nuevos cambios y hoy es necesario proporcionar al ciudadano una educación que tenga en cuenta esta realidad. Estas transformaciones observadas en los procesos de enseñanza y aprendizaje se sitúan en la línea de las teorías constructivistas que ponderan estrategias de aprendizaje que hagan de los alumnos elementos Activos y dinámicos en la construcción del saber (Piaget, Vygotski 2001)

En el estado actual de cosas es normal considerar las nuevas tecnologías como objeto de aprendizaje en sí mismo. Permite que los alumnos se familiaricen con el ordenador y adquieran las competencias necesarias para hacer del mismo un instrumento útil a lo largo de los

estudios, en el mundo del trabajo o en la formación continua (Herín, D 2002).

Se considera que las tecnologías son utilizadas como un medio de aprendizaje cuando es una herramienta al servicio de la formación a distancia, no presencial y del autoaprendizaje o son ejercicios de repetición, cursos en línea a través de Internet, de videoconferencia, CD-ROMs, programas de simulación o de ejercicios, etc. (Meléndez 2000).

Nos centraremos en cuáles son los beneficios que tanto para el alumno como para el profesor tiene la aplicación de las TIC (Busquets, 2000):

- ✓ **Motivación.** El alumno se encontrará más motivado si la materia es atractiva, amena, divertida, si le permite investigar de una forma sencilla utilizando las herramientas TICs o si le permite aprender jugando
- ✓ **Interés.** El docente que impartirá dicha materia se le clasifique como un docente TIC, como un docente que utiliza habitualmente medios informáticos o bien otras herramientas comunicativas para atraer al alumno.
- ✓ **Interactividad.** El alumno puede interactuar, se puede comunicar, puede intercambiar experiencias con otros compañeros. Ello enriquece en gran medida su aprendizaje.
- ✓ **Cooperación**. Las TICs, gracias a la interactividad y posibilidades de comunicación del alumno con sus compañeros o

el equipo docente, también posibilitan la realización de experiencias, trabajos o proyectos en común.

- ✓ **Iniciativa y creatividad**. El desarrollo de la iniciativa del alumno, el desarrollo de su imaginación y el aprendizaje por sí mismo también es una ventaja de estos recursos.
- ✓ **Comunicación**. Es obvio que todo lo anteriormente expuesto se basa en la relación entre alumnos y profesores, una relación muy estrecha en los tradicionales sistemas de enseñanza, pero que permite mayor libertad en los actuales sistemas.
- ✓ **Autonomía**. Existía una mayor dependencia del canal de comunicación que el profesor transmitía al alumno. Ahora, con la llegada de las TICs y la ayuda, sin duda alguna, de Internet sin menospreciar la dirección o guía del profesor el alumno dispone de infinito número de canales y de gran cantidad de información.

Para poder hacer efectivas estas ventajas es importante la formación de los docentes y el desarrollo de habilidades en el uso de las TIC que les permitan transitar junto a los estudiantes el camino hacia la Sociedad de la Información. (Meléndez 2000).

2.2.2. E-learning y B-learning

El E-learning es un modelo de formación a distancia que utiliza Internet como herramienta de aprendizaje. Este modelo permite al

alumno seguir el proceso formativo desde cualquier parte del mundo y a cualquier hora.

E-learning se define como: "Enseñanza a distancia caracterizada por una separación física entre profesorado y alumnado sin excluir encuentros físicos puntuales, entre los que predomina una comunicación de doble vía asíncrona donde se usa preferentemente Internet como medio de comunicación y de distribución del conocimiento, de tal manera que el alumno es el centro de una formación independiente y flexible, al tener que gestionar su propio aprendizaje, generalmente con ayuda de tutores externos". (Germán Ruipérez, 2003).

Según Bernández, se denomina E-learning a "todas aquellas metodologías, estrategias o sistemas de aprendizaje que emplean tecnología digital y/o comunicación mediada por ordenadores para producir, transmitir, distribuir y organizar conocimiento entre individuos, comunidades y organizaciones." (Bernárdez, 2000)

El **B-learning** (formación combinada, del inglés blended learning) consiste en un proceso docente semipresencial; esto significa que un curso dictado en este formato incluirá tanto clases presenciales como actividades de E-learning (Costaguta 2004).

Este modelo de formación hace uso de las ventajas de la formación 100% on-line y la formación presencial, combinándolas en un solo tipo de formación que agiliza la labor tanto del formador como

del estudiante. El diseño instruccional del programa académico para el que se ha decidido adoptar una modalidad B-learning deberá incluir tanto actividades on-line como presenciales, pedagógicamente estructuradas, de modo que se facilite lograr el aprendizaje buscado. (Herín, Sala, Pompidor 2005).

Las ventajas que se suelen atribuir a esta modalidad de aprendizaje son la unión de las dos modalidades que combina:

- Las que se atribuyen al E-learning: la reducción de costes, acarreados habitualmente por el desplazamiento, alojamiento, etc., la eliminación de barreras espaciales y la flexibilidad temporal, ya que para llevar a cabo gran parte de las actividades del curso no es necesario que todos los participantes coincidan en un mismo lugar y tiempo (Zhang C., Zhang S ,2005).
- Las de la formación presencial: interacción física, lo cual tiene una incidencia notable en la motivación de los participantes, facilita el establecimiento de vínculos, y ofrece la posibilidad de realizar actividades algo más complicadas de realizar de manera puramente virtual (Zhang C., Zhang S ,2005).
- Es la combinación de múltiples acercamientos al aprendizaje. El B-learning puede ser logrado a través del uso de recursos virtuales y físicos, "mezclados". Un ejemplo de esto podría ser la combinación de materiales basados en la tecnología y sesiones

cara a cara, juntos para lograr una enseñanza eficaz (W.Klösgen, 2003).

2.2.3. Sistemas E-learning y B-learning en el Perú

Según (Domínguez, J.Rama, C. Rodríguez. 2013), la educación en el Perú, como en el resto del mundo, presenta retos que reclaman cada vez más no solo una educación de calidad sino una actualización en los enfoques, métodos y herramientas, pero sobre todo en la actitud de los protagonistas en él proceso enseñanza-aprendizaje dado los cambios vertiginosos que experimenta la sociedad actual.

En el Perú existen algunas experiencias de educación virtual, la aplicación del E-learning en la modalidad de B-learning y como apoyo y plataforma tecnológica de los cursos presenciales se vienen dando en la mayoría de las Universidades.

En el Perú se vienen realizando campañas de sensibilización respecto a la educación a distancia como un sistema útil y factible, sustentado en los valores como: la pertinencia, la capacidad de innovación en las estrategias utilizadas, la apertura al avance de las TICs como un motor de cambio y desarrollo.

La mayoría de las Universidades en el Perú están usando actualmente plataformas E-learning como apoyo a la docencia presencial, aunque existen universidades que ya están adoptando la modalidad semipresencial o sea el B-learning como alternativa de

formación para pre grado y/o post grado. El avance de las universidades en el Perú viene incorporando gradualmente el uso de las TICs, implementando campus virtuales pilotos mayormente basados en software libre como Moodle y se vienen implementando modelos de aprendizaje basados en TIC en modalidades E-learning y B-learning.

Entre las universidades que ofrecen en la modalidad E-learning y B-learning, por citar a algunas, tenemos la Pontifica Universidad Católica del Perú, Universidad Nacional San Marcos, Alas Peruanas, Católica Los Ángeles de Chimbote, Universidad Inca Garcilaso de la Vega, Universidad José Carlos Mariátegui, y Universidad Peruana Los Andes. Sin embargo son muchas más las universidades con estas modalidades, si contamos a aquellas que vienen implementando estrategias apoyadas en las TICs principalmente para complementar las clases presenciales o para desarrollar diplomados o programas de capacitación.

2.2.4. Los rasgos de aprendizaje en los modelos E-learning y B-learning

La determinación e identificación de los patrones de comportamiento ha sido objeto de análisis por parte de la comunidad científica desde diferentes perspectivas (Anaya y Boticario, 2011). En este sentido, el *Interactive Adverstising Bureau* diferencia tres tipos de usuarios en las redes sociales en función de la actitud que manifiestan con respecto a los demás: los usuarios *Pasivos* son aquéllos que

simplemente buscan, consultan y leen contenidos; los usuarios *Activos* interactúan a nivel de contenidos; y, finalmente, los usuarios *Colaborativos* son generadores de contenido que a su vez se relacionan con otros usuarios. Estos últimos, al proporcionar contenido a la red, son de gran utilidad tanto para los pasivos como para los activos (IAB, 2009). En el contexto de un entorno de enseñanza-aprendizaje esta clasificación también puede ser observada y analizada, en función del grado de participación de los estudiantes en los hilos de debate que se planteen en los foros, bien a iniciativa del profesor o creados por otros estudiantes. A esos tres perfiles de usuarios se podrían añadir un cuarto perfil de usuarios *Inactivos* que realmente optan por no participar de ningún modo o tienen una participación mínima en esos foros de intercambio de experiencias y de debate. En este trabajo se tomará por tanto como referencia esta clasificación de estudiantes según su nivel de participación en los foros, de mayor a menor participación:

1. Usuarios inactivos
2. Usuarios pasivos
3. Usuarios activos
4. Usuarios colaborativos

2.2.5. Gestión de Cursos Virtuales

La educación va evolucionando conjuntamente con la tecnología, un claro ejemplo de ello es la creciente aceptación de las plataformas virtuales en la preparación profesional, el uso constante del internet en

las actividades diarias de las personas hacen que cada vez más usuarios busquen alternativas virtuales para obtener y reforzar conocimientos en el ámbito educativo y profesional.

Para una implementación efectivas de un sistema E-learning o B-learning se requieren programas que puedan llevar a cabo dicha tarea, ahí es donde intervienen las plataformas virtuales, las cuales reciben distintos nombres, tales como "entornos virtuales de enseñanza y aprendizaje" o "entornos de aprendizajes integrados", "ambiente virtual de aprendizaje", "Sistemas de Gestión del Aprendizaje", "Sistema de Gestión de Curso", "Sistema de Gestión de Contenido para el Aprendizaje", "Ambientes de Aprendizaje Gestionado", "Sistema de Apoyo al Aprendizaje", "Plataforma de Aprendizaje"…, pero todos con significados semejantes como podemos ver a continuación en las opiniones de los diversos autores.

Entre las distintas plataformas educativas estandarizadas de uso gratuito disponibles en la red podemos encontrar Claroline, Moodle, Teleduc, Ilias, Ganesha, Fle3, Dokeos.

Entre otras plataformas virtuales no gratuitas que podemos encontrarnos podemos destacar WebCt, eCollege, Angel 5.5, FirstClass, Blackboard

Según (Sebastián Díaz 2009), las plataformas deben poseer unas aplicaciones mínimas, que se pueden agrupar en:

- Herramientas de gestión de contenidos, que permiten al profesor poner a disposición del alumno información en forma de archivos

(que pueden tener distintos formatos: pdf, xls, doc, txt, html…) organizados a través de distintos directorios y carpetas.

- Herramientas de comunicación y colaboración, como foros de debate e intercambio de información, salas de chat, mensajería interna del curso con posibilidad de enviar mensajes individuales y/o grupales.
- Herramientas de seguimiento y evaluación, como cuestionarios editables por el profesor para evaluación del alumno y de autoevaluación para los mismos, tareas, informes de la actividad de cada alumno, plantillas de calificación.

2.2.6. Minería De Datos

La minería de datos (Witten & Frank 2000) se define como el proceso de extraer información válida, previamente desconocida, comprensible y útil de bases de datos almacenados en distintos formatos de gran tamaño y utilizar dicha información para tomar decisiones.

La minería de datos es un componente de la actividad llamada Descubrimiento de Conocimiento en Bases de Datos (KDD, de sus siglas en inglés), es decir en la minería de datos se aplica un algoritmo para extraer patrones de los datos y el KDD es el proceso completo (pre-procesamiento, minería, post-procesamiento). KDD involucra investigación de áreas cómo: aprendizaje computacional, reconocimiento de patrones, bases de datos, estadística, inteligencia artificial, adquisición de conocimiento y visualización de datos.

El descubrimiento de la información oculta es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, se considera como la aplicación de métodos y herramientas, sobre una cantidad grande de datos para lograr la extracción o generación de nuevo conocimiento útil y apoyar la toma de decisiones dentro de un sistema organizacional.

2.2.6.1. Fases de KDD

Se pueden identificar las fases correspondientes al KDD, las entradas y salidas de cada fase, el flujo de la información durante todo el proceso, y el carácter iterativo que tiene. Una implementación del proceso de Descubrimiento de Datos en Bases de Datos puede iniciar desde la parte de selección y transformación.

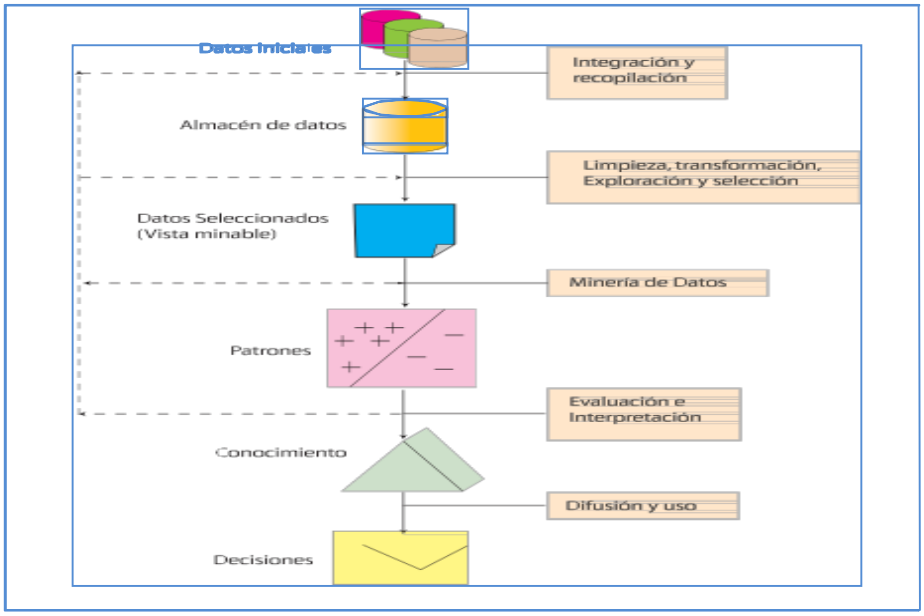


Fig.2.1: Fases particulares del KDD

- **Fase 1 Integración y recopilación**

La integración de los datos, es debida a que éstos provienen de diferentes fuentes, comúnmente en diferentes formatos y se necesita integrarlos a un mismo sistema de almacenamiento que cumpla con las características propias de una estructura ya definida. La recopilación, permite que estos datos sean encontrados, luego de analizar cuáles son los necesarios para poder decir definitivamente que éstos son los que representan las actividades.

- **Fase 2 Limpieza y transformación**

Una vez integrados y recopilados los datos, el siguiente paso (si se hace necesario) consiste en verificar su calidad para que existan las condiciones apropiadas para su análisis, lo cual devuelve al final (una vez aplicada también la fase de exploración y selección) lo que se conoce como vista minable, y que no es más que una estructura de datos con las propiedades necesarias para ser trabajada por los algoritmos de minería de datos.

Gran parte de este proceso de limpieza y transformación es realizado durante las mismas etapas de integración y recopilación mencionadas anteriormente, pero no siempre es necesario tener un almacén de datos.

Las operaciones de transformación de atributos son utilizadas para presentar los datos de una manera idónea para las herramientas de Minería de Datos, haciendo uso de técnicas de discretización, numerización, de reducción y aumento de dimensionalidad, entre otras. Estas operaciones se explican brevemente a continuación:

- ✓ **Numerización:** Se refiere a la conversión del tipo de datos de un valor nominal a un valor numérico. Generalmente, ésta se aplica cuando se pretende utilizar una técnica de Minería de datos
- ✓ **Discretización:** Se refiere a la conversión, en este caso de un valor numérico a un valor nominal ordenado o sin orden.
- ✓ **Aumento de dimensionalidad:** Es creación de nuevos atributos sustituyendo los actuales, o añadiendo otros, para obtener un número mayor de ellos.
- ✓ **Reducción de dimensionalidad:** Es la creación de nuevos atributos sustituyendo los actuales, por un número menor de ellos.
- ✓ **Normalización de rango*:*** Éste se refiere a normalizar el rango de un valor numérico para que los atributos estén en función de la misma medida.

- **Fase 3 Exploración y selección**

Se debe realizar un reconocimiento o análisis exploratorio de los datos que se hallen en los Almacenes de Datos, con el objetivo de conocerlos mejor y poder definir cuales datos seleccionar y determinar las tareas a realizar sobre los mismos.

Aquí se ven técnicas simples del análisis exploratorio de los datos, técnicas de visualización previa, de agrupamiento exploratorio, técnicas de selección horizontal y vertical, interfaces gráficas y técnicas de consulta y agregación.

- **Fase 4 Minería de datos**

En la secuencia normal de las fases que se ejecutan en el proceso de descubrimiento de conocimiento en bases de datos, continua la fase de Minería de Datos, pues de la anterior ya se obtuvo como resultado una vista minable lista para que le sean aplicados los algoritmos correspondientes y luego de esto obtener como resultado un modelo que será evaluado en la siguiente fase.

En esta fase se aplican procesos a los datos para extraer patrones y obtener posteriormente conocimiento útil y apoyar la toma de decisiones en las empresas, todo esto por medio de la aplicación de diferentes técnicas. Las principales técnicas de minería de datos que pueden ser aplicadas son las de clasificación

supervisada, clasificación no supervisada o clustering, y selección de atributos.

- **Fase 5 Evaluación e interpretación**

La validación y verificación de los modelos que se plantearon en la fase anterior, y así poder reconocer en qué nivel de madurez o de refinamiento se encuentra el modelo y si es el más adecuado para resolver el problema planteado.

En general se proponen tres métodos, los cuales sirven como una aproximación a las diferentes técnicas utilizadas para partir los datos en los conjuntos de entrenamiento y de prueba.

- ✓ **Validación simple:** reserva un porcentaje de la base de datos como conjunto de prueba, y no lo usa para construir el modelo.

- ✓ **Validación cruzada con n pliegues:** Este método generalmente se aplica cuando no se cuenta con la cantidad suficiente de datos como para construir el modelo a partir del conjunto de entrenamiento.

- ✓ **Evaluación basada en coste:** Se evalúa el coste de los errores cometidos por un modelo, y no el modelo que cometa menor número de errores.

- **Fase 6 Difusión uso y monitorización**

Una vez construido y validado el modelo puede usarse principalmente con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos.

2.2.6.2. Técnicas de la Minería de Datos

Las técnicas se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones. Se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Tobías S. 2008). La siguiente tabla recoge algunas de las técnicas más habituales.

Arboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento ("Clustering")
Series temporales	Reglas de asociación
	Patrones secuenciales

Tabla2.1: Algunas de las técnicas de minería

2.2.7. Minería De Datos Educativos (EDM)

El desarrollo tradicional de los cursos E-learning es una actividad ardua en el que el profesor del curso tiene que elegir el contenido que se mostrará, decidir sobre la estructura de los contenidos, y determinar los elementos de contenido más apropiado para cada tipo de usuario potencial del curso. (Romero, Ventura, & García 2009). La minería de datos en este caso pueda ayudar al profesor a identificar aquellos aspectos de su curso que puedan ser susceptibles de mejora, a conocer mejor las dificultades que sus estudiantes se encuentran a la hora de desenvolverse dentro del correspondiente curso virtual, a integrar sistemas inteligentes de asignación de tareas.

El sitio web http://www.educationaldatamining.org/EDM/ de la comunidad de Educational Data Mining (EDM), define la minería de datos educativos de la siguiente manera: "La minería de datos en la educación es una disciplina emergente, cuyo interés radica en la elaboración de métodos para explorar los tipos de información que proceden de los centros educativos y el uso de los métodos para comprender mejor a los estudiantes y la manera en que aprenden".

Según (Corso & Alfaro, 2010), el Data Mining tiene como objetivo reunir los beneficios de varias áreas como la estadística, inteligencia artificial, las bases de datos y el preprocesamiento masivo, usando las bases de datos como materia prima.

Existen diversos contextos donde se podría implementar EDM, (Baker & Yacef, 2009) manifiestan la existencia de cuatro áreas claves:

- ✓ **La primera:** En los últimos años los investigadores han utilizado EDM para deducir el desenvolvimiento del estudiante dentro del sistema y lo aburrido o frustrado que podría sentirse, han podido ampliar además el modelo de estudiante para la determinación de posibles fracasos o falta de retentiva del alumno.
- ✓ **La segunda:** Para el descubrimiento o mejora de la estructura de los modelos de conocimiento de dominio, algunos investigadores han sido capaces de desarrollar enfoques automatizados que se pueden descubrir modelos precisos de estructura de dominio, directamente de los datos.
- ✓ **La tercera**: Un tercer aspecto clave de la aplicación de métodos de EDM ha sido en el estudio pedagógico de apoyo (tanto en software para el aprendizaje y el aprendizaje en otros dominios, como los comportamientos), para descubrir qué tipos de apoyo pedagógico son más eficaces, ya sea de forma general o por grupos o perfiles de estudiantes o en situaciones diferentes.
- ✓ **La Cuarta:** La búsqueda empírica de pruebas para perfeccionar y ampliar las teorías educativas y fenómenos

educativos conocidos, para una comprensión más profunda de los factores clave que afectan el aprendizaje.

Según (Romero & Ventura 2010), una consideración inicial, parece implicar sólo dos grandes grupos de usuarios interesados: los alumnos y los instructores. En realidad hay más grupos que participan con muchos más objetivos, como puede verse en la figura 2.2.

Usuarios/actores	Objetivos de uso- Data Mining
Alumnos/Estudiantes/Pupilos	Personalizar el e-learning, realizando una recomendación de actividades y tareas, forjando un aprendizaje basado en experiencia.
Educadores/Profesores/ Instructores/tutores	Obtener una retroalimentación acerca de la enseñanza, analizando el comportamiento del estudiante detectando que estudiantes necesitan un mayor soporte y que errores son los más comunes que se puede llegar a tener personalizando, adaptando los cursos para un mejor aprendizaje y usabilidad.
Desarrolladores de Cursos/Investigadores Educativos	Para la evaluación y mantenimiento de cursos, valorando la estructura de contenido de los cursos, comparando técnicas de minería de datos con el fin de recomendar las más útiles para cada tarea.
Organizadores/Proveedores de aprendizaje/Universidades/ Empresas de formación privada	Para la toma de decisiones en instituciones de nivel superior. Encontrando la mejor relación costo/eficiencia. Seleccionando los candidatos más calificados para la admisión en sus universidades.
Administradores/Administradores de centro educativo/Administradores de red/Administradores de Sistema	Para desarrollar la mejor manera de organizar los recursos institucionales humanos y materiales y su oferta educativa, para establecer parámetros de eficiencia del sitio, determinando el enfoque y eficiencia de la educación a distancia.

Fig. 2.2: EDM Usuarios/Objetivos

Para facilitar todas las tareas propias de la EDM, se necesitan métodos de análisis de datos y herramientas para observar el comportamiento de los estudiantes y maestros, para ayudar en la detección de posibles errores, deficiencias y posibles mejoras. El análisis de datos tradicional en el E-learning es la hipótesis que más impulso tiene en el sentido de que el usuario parte de una pregunta y explora los datos para confirmar su intuición.

En la minería de datos se extraen automáticamente nueva información y conocimiento a partir de los datos que se obtengan para el análisis en lugar de basarse en la investigación o impresión humana. La EDM construye modelos de análisis que descubren patrones y tendencias interesantes de información sobre el uso del estudiante que pueden ser utilizadas por el profesor para mejorar el aprendizaje del estudiante y el mantenimiento del sistema.

Según (Rodríguez Anaya, 2009) las principales aplicaciones de la EDM en un entorno de enseñanza apoyado en la Web son: obtener una mayor realimentación de la enseñanza, conocer más sobre cómo los estudiantes aprenden, evaluar a los estudiantes por sus patrones de navegación, reestructurar los contenidos el sitio web para personalizar los cursos, clasificar a los estudiantes en clústeres, etc.

Se han realizado distintos experimentos para medir o identificar la colaboración que se realizaba entre usuarios de un sistema de E-learning con diferentes orientaciones:

- ✓ ***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.

- ✓ ***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 aprendiz, ayudar a los estudiantes dando sugerencias y atajos, recomendar caminos más cortos y personalizados, etc.

a.- Revisión de la investigación actual en este campo.

Entre las investigaciones que aplican técnicas de minería de datos en educación se puede destacar según Baker, R.S.2014) Educational Data Mining: An Advance for Intelligent Systems in Education. IEEE Inteligent Systems. (A. Cobo, R. Rocha, C. Rodríguez, 2014) Evaluación de la interactividad de los alumnos en entornos de aprendizaje virtuales utilizando un enfoque multicriterio y minería de datos. (Palmer, 2013) datos de los estudiantes usados almacenados en los sistemas institucionales para predecir el rendimiento académico de los estudiantes. (Gil, Herrera et al. 2011) centrado en la predicción del rendimiento académico utilizando áspera la teoría de conjuntos.

(C. Romero, S. Ventura 2013) Data Mining and Knowledge Discovery in Education. (Cobo, A; Rocha, R; Álvarez, 2012) Análisis y Evaluación del Rendimiento Académico y Patrones de Comportamiento de los estudiantes en un entorno Interactivos B-learning Utilizando Técnicas De Minería De Datos. (Bodea et al.

2012) presentó un recomendador educativa motor para aumentar los valores formativos de la E-learning.

(Anaya, A; Boticario J.G. 2011) Aplicación de técnicas de aprendizaje automático que analiza las interacciones de los estudiantes y mejorar el proceso de colaboración (García y Zorrilla, 2011) los utilizaron para analizar el comportamiento de los estudiantes en los cursos on-line y sus procedimientos de navegación. (Johnson et al. 2011) utilizado a analizar los datos obtenidos de los registros de actividad (logs) generada por los sistemas de tutoría automáticas.

(C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura 2010) Web Usage Mining for Predicting Final Marks of Students that Use Moodle Courses. Computer Applications in Engineering Education journal.

b. - Eventos relacionados con EDM en los últimos años

- The 8th International Conference on Educational Data Mining EDM 2015. http://www.educationaldatamining.org/EDM2015/
- The 7th International Conference on Educational Data Mining ED 2014 Institute of Education, London, UK. http://www.educationaldatamining.org/EDM2014/

- Predicting School Failure and Dropout by Using Data Mining Techniques IEEE Journal of Latin-American Learning Technologies, 2013.
- Web Usage Mining for Predicting Final Marks of MOODLE Students. Computer Applications in Engineering Education, the Internet and Higher Education Journal (Special Issue on Web Mining and Higher Education 2013).
- Multi-objective ant programming for mining classification rules. Fifteen European Conference on Genetic Programming (EuroGP 2012), Malaga, España, Abril de 2012.
- 4th International Conference on Educational Data Mining (EDM 2011) Eindhoven,
- 3th International Conference on Educational Data Mining (EDM 2010) Eindhoven, Pittsburgh 2010,
- Workshop on Educational Data Mining, at the 7th IEEE International Conference on Advanced Learning Technologies. Niigata, Japan. July 18-20, 2007.
- Track on Educational Data Mining, at the Workshop on Data Mining for User Modeling, at the 11th International Conference on User Modeling. Corfu, Greece. June 25, 2007.

c.- Aplicaciones de la minería de datos en educación

Las principales aplicaciones de las técnicas de minería de datos en educación (Romero y Ventura; 2010) han elaboraron una lista de

aplicaciones, en el que describen los estudios más relevantes agrupados en función de diferentes criterios:

- Predicción del rendimiento académico
  - Predecir qué probabilidad tienen los estudiantes de cometer errores en un determinado problema.
  - Las reglas de asociación guían el proceso de búsqueda para encontrar modelos de transferencia que predicen el éxito de los estudiantes.
- Sistemas de tutorización inteligente
  - Recomendación de tareas
  - Entrenamiento a un agente recomendador y construir un modelo del comportamiento de acceso o asociaciones entre actividades de aprendizaje en línea.
  - Descubrir patrones que reflejan comportamientos de los usuarios para dar soporte a la tutoría en comunidades de aprendizaje virtual
- Identificación de debilidades de los sistemas de tutoría
  - Un buen estudiante proporciona información útil sobre la dificultad en las tareas de aprendizaje y la transferencia de aprendizaje entre determinados problemas.
- Sistemas de seguimiento de conocimiento
  - Descubrimiento de relaciones o asociaciones entre distintas páginas Web visitadas

- Analizar secuencias de páginas visitadas durante una sesión o en distintas sesiones de un mismo usuario.

- Sistemas de modelado de estudiantes
  - Utilizado para identificar diferentes tipos de estudiantes, sus características, comportamientos de interacción y relación con el aprendizaje.
  - Agrupar a los usuarios por su comportamiento de navegación.
  - Mejorar las estrategias en los sistemas instruccionales.
- Sistemas de detección de irregularidades
  - Detección en línea de irregularidades en perfiles de procesos de aprendizaje a partir de los tiempos de respuesta
  - Identificación automática de plagio.
- Sistemas clasificadores de contenidos.
  - Agrupar a las páginas por su contenido o tipo o acceso.
  - Generación automática de metadatos.
  - Construcción automática de libros electrónicos a partir de una jerarquía de temas especificados por el usuario.
  - Utilización conjunta de técnicas de minería de secuencia de patrones para descubrir nuevos conceptos en las páginas web.

2.2.8. Proceso de Análisis Jerárquico (AHP)

Dentro del Análisis de Decisiones Multicriterio una de las técnicas más populares es el método denominado Proceso Analítico Jerárquico, AHP (Analytical Hierarchy Process) propuesto por Saaty (Thomas Saaty, 1980). Esta metodología se basa en construir un modelo jerárquico que represente el problema objeto de estudio, mediante criterios y alternativas planteadas inicialmente, luego poder deducir, cual o cuales son las mejores alternativas y tomar una decisión final óptima. Con este proceso analítico jerárquico se pretende dividir una decisión compleja en un conjunto de decisiones simples, facilitando la comprensión y solución del problema propuesto.

El AHP es una metodología para estructurar, medir y sintetizar. Ha sido aplicada ampliamente en la solución de una gran variedad de problemas, es un método matemático creado para evaluar alternativas cuando se tienen en consideración varios criterios y está basado en el principio que la experiencia y el conocimiento de los actores son tan importantes como los datos utilizados en el proceso.

El AHP se fundamenta en:

- La estructuración del modelo jerárquico (representación del problema mediante identificación de meta, criterios, subcriterios y alternativas).
- Priorización de los elementos del modelo jerárquico.
- Comparaciones binarias entre los elementos.
- Evaluación de los elementos mediante asignación de "pesos".

- Ranking de las alternativas de acuerdo con los pesos dados.
- Síntesis.
- Análisis de Sensibilidad.

El AHP utiliza comparaciones entre pares de elementos, construyendo matrices a partir de estas comparaciones, y usando elementos del álgebra matricial para establecer prioridades entre los elementos de un nivel, con respecto a un elemento del nivel inmediatamente superior. Cuando las prioridades de los elementos en cada nivel se tienen definidas, se agregan para obtener las prioridades globales frente al objetivo principal.

La metodología AHP se basa en tres principios básicos. El primero el de descomposición, que permite estructurar un problema complejo en subproblemas jerárquicos con dependencias de acuerdo con el nivel de descomposición en el que se encuentren. El principio de juicio comparativo se basa en realizar combinaciones en parejas de todos los elementos de un sub-criterio con respecto al criterio principal, por ello se habla de comparaciones biunívocas. En las comparaciones dos a dos se suele utilizar la conocida como escala de Saaty y que se recoge en la Tabla 2.2.

Escala	Definición	Explicación
1	Igualmente preferida	Los dos criterios contribuyen igual al objetivo
2	Moderadamente preferida	La experiencia y el juicio favorecen un poco a un criterio frente al otro
5	Fuertemente preferida	La experiencia y el juicio favorecen fuertemente a un criterio frente al otro
7	Muy fuertemente preferida	Un criterio es favorecido muy fuertemente sobre el otro.
9	Extremadamente preferida	La evidencia favorece en la más alta medida a un factor frente al otro
2,4,6,8	Valores intermedios	Cuando se necesita un compromiso de las partes entre valores adyacente
Recíprocos	aij =1/aij	Hipótesis del método

Tabla 2.2: Escalas de comparación de Saaty

Finalmente el principio de composición jerárquica o síntesis de prioridades permite producir prioridades globales a través de las multiplicaciones de las prioridades locales, es decir, que una vez se tienen soluciones locales, se agregan para obtener la solución general que se está buscando. En el cálculo de las diferentes prioridades pueden utilizarse diferentes técnicas algebraicas, una de las más comunes es la consideración del autovector normalizado asociado al autovalor dominante de la matriz de comparaciones dos a dos.

2.2.9. Herramienta Weka

WEKA (http://www.cs.waikato.ac.nz/ml/weka/) es una colección de algoritmos de aprendizaje para tareas de Minería de Datos. Se trata de una herramienta totalmente desarrollada en el lenguaje de programación JAVA, que contiene una interfaz gráfica para la aplicación de los algoritmos

directamente al conjunto de datos de origen, así como la posibilidad de aplicar los algoritmos desde un programa JAVA independiente.

WEKA según (Isa 2003) es un software libre, está establecido por una serie de paquetes de código abierto con diferentes técnicas de reprocesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos. Con objeto de facilitar su uso por un mayor número de usuarios, WEKA además incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas.

Algunas de las principales características de esta herramienta son:

- Interacción con datos de origen: Puede recibir los datos de origen directamente desde una base de datos, utilizando protocolo JDBC, el cual fue el método utilizado para la elaboración de las pruebas de este proyecto.

- Interfaz gráfica: provee herramientas de visualización, preprocesamiento y métodos para la aplicación de los algoritmos de Minería de Datos.

- Preprocesamiento: son llamados “filtros” y permiten realizar actividades como discretización, normalización, selección de atributos, transformación y combinación de los mismos.
- Algoritmos de clasificación clustering, asociativa: en WEKA, que implementan una gran variedad de algoritmos de Minería de Datos

2.2.10. Herramienta Expert Choice

Expert Choice (www.expertchoice.com) es un software de apoyo para la toma de decisiones, está basado en el Proceso Jerárquico Analítico (AHP, Analytic Hierarchy Process) descrito anteriormente.

Expert Choice asiste a los decisores organizando la información relacionada a la complejidad del problema en un modelo jerárquico consistente de un objetivo, escenarios posibles, criterios y alternativas. Usando el método, único de Expert Choice, de comparación par a par, se puede evaluar la importancia de los criterios, las preferencias de las alternativas, y las probabilidades de los escenarios y sintetizar sus comparaciones para llegar a la mejor decisión.

Expert Choice ha sido usado exitosamente en una variedad de aplicaciones incluyendo Priorización y Evaluación de Proyectos, Planeamiento estratégico, Análisis de costos, Negociación y solución de conflictos, Evaluación de los planes de inversión, Asignación de recursos, Administración de Calidad total, Evaluación de fusiones y adquisiciones, Contratación, evaluación y promoción de empleados, Satisfacción del cliente.

2.3. DEFINICIÓN DE TERMINOS

En particular, a efectos de esta tesis resulta interesante analizar algunos métodos comunes de clustering, pues ésta es la técnica de minería de datos que se utilizará en este trabajo.

- **Técnicas de Clustering**

Clustering es una de las tareas más utilizadas en el proceso de minería de datos. El objetivo de esta técnica es descubrir grupos o clústeres dentro de los datos e identificar interesantes distribuciones y patrones en ellos. El objetivo del clustering es organizar una muestra de casos en consideración en grupos de forma tal que el grado de asociación es alto entre los miembros del mismo grupo (clúster) y bajo entre los miembros de diferentes grupos. En definitiva, se busca que los elementos asignados a un mismo grupo se parezcan entre sí lo máximo posible y que al mismo tiempo se diferencien al máximo de los elementos de los otros grupos. Para analizar esas similitudes o diferencias se necesita contar con las correspondientes métricas o medida.

Las medidas de similitud representan la similitud como la proximidad de las observaciones respecto a las otras. Las medidas de distancia utilizadas para el agrupamiento pueden ser también muy diversas; entre las más usadas se encuentran:

- **Distancia euclidiana:** definida como la longitud del segmento que une dos puntos en el espacio, se deduce a partir del Teorema de Pitágoras.

$$Eucl(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

- **Distancia de Manhattan:** esta hace referencia a recorrer un camino más corto zigzagueando. El nombre viene precisamente como alusión del arreglo en rejilla de la mayor parte de las calles en la isla de Manhattan, propuesta por Hermann Minkowski (IOS, 2006).

$$Manh(x,y) = \sum_{i=1}^{d} | x_i - y_i |$$

- **Distancia de Mahalanobis:** fue introducida por Mahalanobis en 1936. Es una distancia más robusta que utiliza la matriz de covarianzas para incorporar la dependencia entre las dos variables. La distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad y con matriz de covarianza Σ se define como:

$$Mahalanobis(x,y) = \sqrt{(x-y)Cov(D)^{-1}(x-y)}$$

Los algoritmos de clústering se dividen en dos grandes grupos: los Agrupamientos Numéricos y los Simbólicos. Entre los primeros

están algoritmos como el de las K-Medias o el algoritmo EM (Expectation Maximation). Entre los algoritmos simbólicos se podrían citar los algoritmos de clustering Jerárquico.

En este trabajo se hará explícitamente uso de los dos algoritmos numéricos citados:

- **Algoritmo K-Medias:** Se trata de un algoritmo clasificado como método de particionado y recolocación. El nombre le viene porque representa cada uno de los clústeres por la media (o media ponderada) de sus puntos, es decir, por su centroide.

El proceso con el que se desenvuelve el algoritmo según (Molina & García, 2004) consiste en:

1. Elegir k elementos (instancias de los datos) que actúan como semillas (k sería el número de clústeres a crear y debe ser dado de antemano).
2. Para cada elemento del conjunto de datos seleccionar la semilla más próxima y añadirlo a la clase o clúster correspondiente.
3. Calcular el centroide de cada clase, que pasan a ser las nuevas semillas
4. Si no se llega a un criterio de convergencia (por ejemplo, dos iteraciones consecutivas que no cambian las clasificaciones de los ejemplos), volver al paso 2.

Algoritmo EM (Expectation Maximation): el algoritmo EM asigna a cada instancia una distribución de probabilidad de pertenencia a cada clúster. El algoritmo puede decidir cuántos clústeres crear a partir de una validación cruzada, o también se le puede especificar a priori cuantos debe generar. El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir, cómo de bien encajan los datos sobre la distribución que los representa. Este valor de bondad se conoce como el verisimilitud de los datos. Se conforma principalmente de dos pasos:

- **(Expectation)** para cada instancia calcular la probabilidad de pertenecer a un clúster.
- **(Maximization)** estimar los parámetros que caracterizan las distribuciones a partir de las nuevas probabilidades (hacer más probable esa observación).

El hecho de que el algoritmo EM pueda decidir de forma automática cuantos grupos crear es una de sus principales ventajas con respecto a otros como el de las k-medias en las que ese número debe ser prefijado.

III METODOLOGIA

Introducción

Dada la orientación de esta investigación, cuyo objetivo es colectar datos de los modelos E-learning y B-learning para estudiar aspectos como la frecuencia y periodicidad de accesos, el seguimiento de las actividades, la participación en los foros, haciendo uso de técnicas de minería de datos, los pasos que se proponen son los siguientes: preparación de los datos, aplicación de las técnicas seleccionadas, evaluación del modelo obtenido, y uso del modelo sobre un caso concreto. Para ello se comenzará con la etapa de pre-procesamiento que incluye la selección y limpieza de datos, luego se emplearán los algoritmos K-MEANS y EM como instrumento para la determinación de patrones de comportamiento similares entre los estudiantes. Para la identificación de perfiles asociados a cada uno de los patrones se hará uso de un indicador de actividad global obtenido mediante la metodología AHP. Todo ello se hará sobre dos experiencias concretas, una que responde a un modelo B-learning y otra que responde a un modelo totalmente on-line.

3.1. TIPO Y DISEÑO DE INVESTIGACIÓN

3.1.1. De acuerdo a la orientación

Aplicada

3.1.2. De acuerdo a la técnica de contrastación

Descriptiva – Explicativa

3.2. PLAN DE RECOLECCIÓN DE LA INFORMACIÓN

3.2.1. Población

Se trata de identificar patrones de comportamiento de estudiantes universitarios que participan en procesos de enseñanza virtual sobre plataformas E-learning, tanto en modelos mixtos como totalmente on-line. La población objeto de estudio es por tanto la formada por usuarios matriculados en cursos virtuales.

3.2.2. Muestra

Se trabajará con una muestra de estudiantes matriculados en dos asignaturas diferentes de la Universidad de Cantabria (España). Dicha universidad es una universidad presencial pero que cuenta con sistemas de enseñanza virtual que se utilizan como apoyo para diferentes asignaturas presenciales pero también para asignaturas impartidas íntegramente on-line.

En este caso se consideraron dos muestras de estudiantes:

1. Los 343 estudiantes inscritos en una asignatura sobre "Teoría de la Optimización" de primer curso de dos titulaciones: Economía y Administración y Dirección de Empresas. Esta asignatura respondía a un modelo B-learning.
2. Los 57 estudiantes de primer curso de diferentes carreras inscritos en una asignatura 100% virtual sobre "Herramientas de Trabajo Colaborativo". Esta asignatura respondía a un modelo E-learning.

A continuación se recogen las principales características de la organización de las dos asignaturas analizadas:

	Teoría de la Optimización	Herramientas de apoyo al trabajo colaborativo
Modelo de enseñanza	B-learning	E-learning
Número de estudiantes	343	57
Ciclo	1º	1º
Duración de la asignatura	4 meses	2 meses
Titulaciones	Economía/Adm. Empresas	Todas las de la UC
Actividades realizadas presencialmente	Clases teóricas y prácticas. Participación en prácticas de Laboratorio. Evaluaciones presenciales	Ninguna
Actividades realizadas virtualmente	Participación en foros, acceso a materiales	Participación en foros, acceso a materiales

	docentes, tests de autoevaluación, entrega de tareas	docentes, tests de autoevaluación, entrega de tareas
Número de foros disponibles	4 foros temáticos más un foro general y otro de noticias	4 foros temáticos más un foro general y otro de noticias
Número de tests on-line a realizar	4	4
Plataforma virtual	Moodle	Moodle

Como puede apreciarse ambas asignaturas responden a una estructuración muy similar: organizadas en 4 bloques temáticos, con un foro de debate y un test asociado a cada bloque, y con tareas que debían ser entregadas a través de la plataforma. La diferencia esencial es que en la primera se realizaban también actividades docentes presenciales mientras que en la segunda todo era on-line.

3.3 INSTRUMENTO(s) DE RECOLECION DE INFORMACION

La recolección de información de datos supuso realmente dos fases diferenciadas:

1. Utilización de las herramientas de generación de informes de actividad de la plataforma Moodle de la Universidad de Cantabria para extraer registros de acceso de los estudiantes a los cursos virtuales de la asignatura. A partir de esos registros se generarán los indicadores individuales para cada estudiante que serán

la base para la aplicación de las técnicas de minería de datos.

2. Consulta a los profesores responsables de las asignaturas para generar matrices de comparación entre criterios a la hora de valorar la importancia de la participación de los estudiantes en las actividades con soporte virtual. Estas matrices serán la base para determinar, usando la metodología AHP, los pesos de cada indicador para el cómputo de un grado de interactividad global de cada estudiante.

3.4. PLAN DE PROCESAMIENTO Y ANALISIS ESTADISTICO DE INFORMACION

3.4.1 Selección de datos

La figura 3.1 muestra la herramienta de generación de informes de actividad de la plataforma Moodle de la Universidad de Cantabria. A través de esa herramienta se puede acceder a registros de acceso a la plataforma de usuarios individuales o del conjunto de estudiantes. También es posibles filtrar esos registros por fechas o por tipo de actividad realizada. En este caso se obtuvieron, para las dos asignaturas consideradas, todos los registros de actividad de todos los estudiantes inscritos en cada una de las asignaturas.

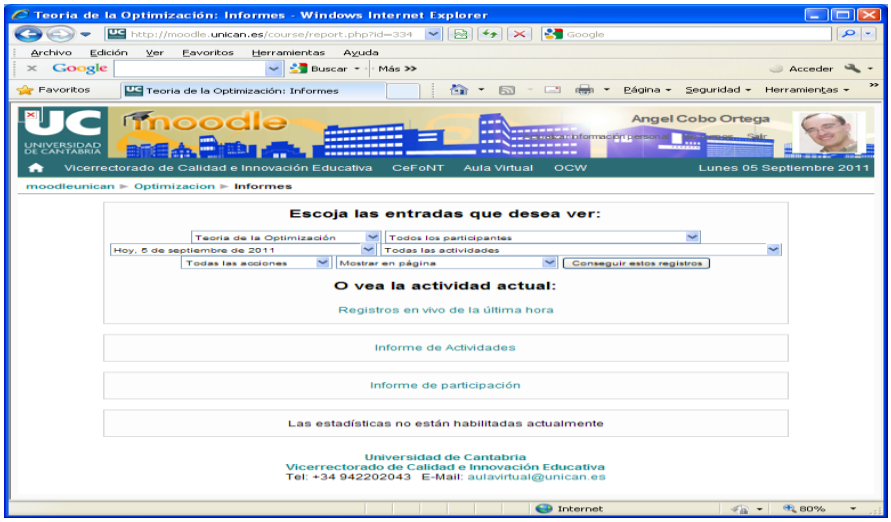


Fig. 3.1: Curso en la Plataforma Moodle

El tipo de informe que se genera con esta herramienta de la plataforma Moodle es el que se muestra la figura 3.2. Como puede apreciarse, quedan registrados todos los accesos señalando para cada uno de ellos la fecha y hora del acceso, la dirección IP de la máquina desde la que se produce la conexión, el tipo de actividad realizada de acuerdo a una clasificación del propio Moodle, y el elemento del curso (foro, elemento de contenido, test,…) sobre el que se ha realizado la actividad.

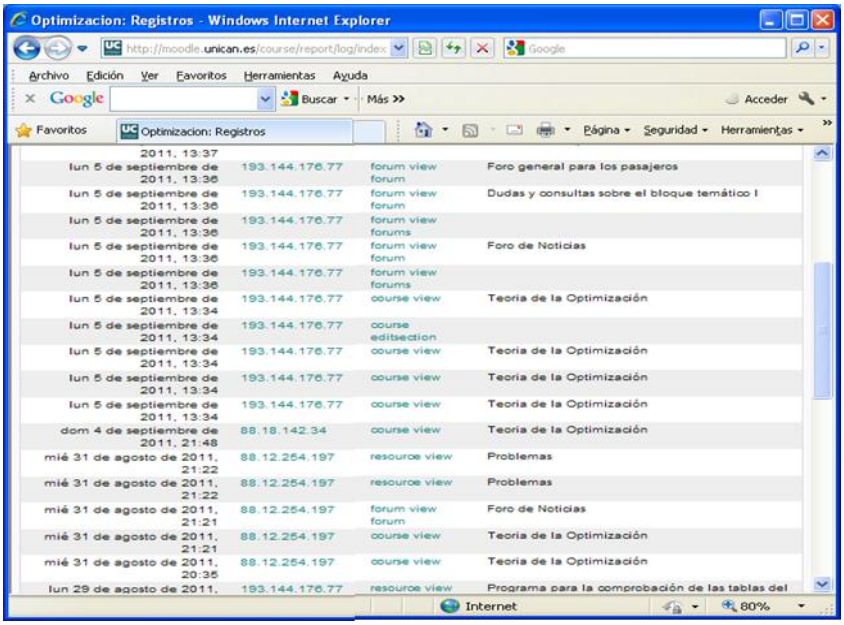


Fig. 3.2: Registros utilizados para cada estudiante

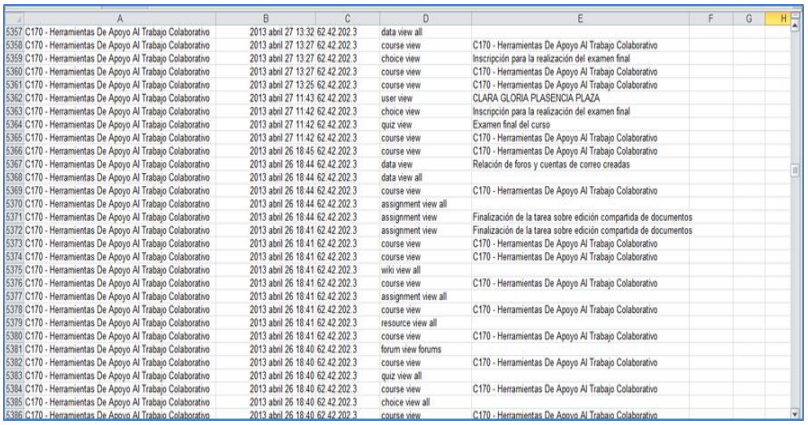
Respecto a la tipología de actividades que define el propio Moodle, éstas son las que se recogen en la tabla 3.1. En este caso se ha añadido a la tabla un código y un nombre breve de la actividad para facilitar la posterior identificación de cada una. Como puede preciarse, queda registrado en la plataforma prácticamente cualquier hecho que el usuario pueda realizar sobre un elemento de un curso virtual. En particular interesa especialmente observar que hay actividades relacionadas con los foros de debate, consultas on-line, tests de evaluación, tareas y recursos o materiales docentes.

cod	Actividad	tipo de acción	Descripción
1	choice1	choice choose	Elección de una opción en una consulta
2	choice2	choice choose again	Cambio de la elección realizada
3	choice3	choice view	Vista de una consulta
4	foro1	forum add discussion	Escriben un mensaje en un foro
5	foro2	forum add post	Respuesta a un mensaje previo
6	foro3	forum delete discussion	Borrar un mensaje de un foro
7	foro4	forum delete post	Borrar una respuesta de un foro
8	foro5	forum subscribe	Suscripción a un foro
9	foro6	forum unsubscribe	Eliminar la suscripción a un foro
10	foro7	forum update post	Modificación de un mensaje enviado a un foro
11	foro8	forum view fórum	Entrada a un foro
12	foro9	forum view discussion	Lectura de un mensaje de un foro
13	quiz1	quiz attempt	Inicio de un test de evaluación
14	quiz2	quiz close attempt	Finalización de un test de evaluación
15	quiz3	quiz continue attemp	Continuación de un intento de un test
16	quiz4	quiz review	Revisión de los resultados de un test
17	quiz5	quiz view	Vista de las condiciones de realización
18	Resource	resource view	Vista de materiales publicados en el curso
19	trabajo1	assignment upload	Carga de un trabajo o tarea
20	trabajo2	assignment view	Vista de las condiciones de un trabajo

Tabla 3.1: Descripción de los tipos de actividades que quedan registrados en la plataforma Moodle.

3.4.2. Preprocesado de datos

Lo primero que se debe hacer para poder realizar un análisis de datos, es obviamente contar con ellos en un formato que permita el análisis automatizado. Para tal efecto, lo primero que se realizó fue la exportación de los informes de actividad a hojas Excel. En la figura 3.3 se muestra la hoja correspondiente a las actividades totales en la segunda de las asignaturas, la impartida totalmente en modalidad on-line.



	A	B	C	D	E	F	G	H
5357	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 13:32	62	42	202 3	data view all		
5358	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 13:27	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5359	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 13:27	62	42	202 3	choice view	Inscripción para la realización del examen final	
5360	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 13:27	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5361	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 13:26	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5362	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 11:43	62	42	202 3	user view	CLARA GLORIA PLASENCIA PLAZA	
5363	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 11:42	62	42	202 3	choice view	Inscripción para la realización del examen final	
5364	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 11:42	62	42	202 3	quiz view	Examen final del curso	
5365	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 27 11:42	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5366	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:45	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5367	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:44	62	42	202 3	data view	Relación de foros y cuentas de correo creadas	
5368	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:44	62	42	202 3	data view all		
5369	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:44	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5370	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:44	62	42	202 3	assignment view all		
5371	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:44	62	42	202 3	assignment view	Finalización de la tarea sobre edición compartida de documentos	
5372	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	assignment view	Finalización de la tarea sobre edición compartida de documentos	
5373	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5374	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5375	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	wiki view all		
5376	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5377	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	assignment view all		
5378	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5379	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	resource view all		
5380	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:41	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5381	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	forum view forums		
5382	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5383	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	quiz view all		
5384	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	
5385	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	choice view all		
5386	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	2013 abril 26 18:40	62	42	202 3	course view	C170 - Herramientas De Apoyo Al Trabajo Colaborativo	

Figura 3.3: Base de datos y cargada en Excel

Para complementar la investigación y para poder cumplir con el objetivo propuesto que es el de identificar los rasgos de comportamientos de los estudiantes en el aprendizaje colaborativo apoyado en los sistemas E-learning y B-learning, se calcularon el número total de registros de actividad asociados a cada estudiante y a cada tipo de actividad. Así por ejemplo, para cada estudiante se pudo calcular el número total de acceso a un foro para leer un mensaje, el número total de respuestas dadas a mensajes de compañeros, el número total de nuevos hilos de debate abiertos en los foros, el número total de veces que accedieron a un elemento de contenido del curso, el número total de accesos a las herramientas de entrega de tareas.

3.4.3. Integración de datos:

A través de Moodle se obtuvieron para cada estudiante datos relativos a su actividad en la parte no presencial de la asignatura, que en el caso de la segunda asignatura eran todos los indicadores disponibles.

En concreto, para cada estudiante de cada asignatura, se obtuvieron los siguientes atributos:

- ✓ Actividad en el foro de noticias (FN), creado por el equipo docente para informar sobre aspectos importantes de la asignatura. En él los estudiantes únicamente podían realizar operaciones de lectura, por lo que se tiene un único atributo a partir del número de mensajes leídos (FNL).
- ✓ Actividad en el foro General (FG), abierto a todo tipo de participación. Se consideran 3 atributos para este foro: mensajes leídos (FGL), respuestas a mensajes de otros usuarios (FGR) y planteamiento de nuevos hilos de debate (FGN).
- ✓ Actividad en los cuatro foros temáticos, uno por cada bloque de la asignatura (F1, F2, F3, F4): abierto igualmente a todo tipo de participación y con tres atributos cada uno por usuario (FxL, FxR, FxN).
- ✓ Resultados en los tests de autoevaluación en cada uno de los bloques temáticos (atributos T1, T2, T3 y T4).
- ✓ Evaluación de un trabajo práctico a distancia (TR), el trabajo era realizado de manera autónoma y entregado a través de la plataforma virtual.

En definitiva, para cada estudiante se dispone de atributos que permiten medir su rendimiento en las actividades y foros, como se muestra en las tablas 3.3, 3.4.

Atributo	Origen	Descripción
Foro general (FG)	FL FR FN	Mensajes leídos Respuestas ofrecidas a otros Nuevos mensajes de debate
Foro Temáticos(Fx)	F1 F2 F3 F4	F1L, F2L, F3L, F4L F1R, F2R, F3R,F4R F1N,F2N,F3N,F4N
Test de autoevaluación	T1 T2 T3 T4	Test autoevaluación 1 Test autoevaluación 2 Test autoevaluación 3 Test autoevaluación 4
Trabajos	TR	Trabajo

Tabla 3.3: Atributos generados para cada estudiante

Con objeto de poder tener información que permita contrastar si la actividad en el curso virtual de un estudiante guarda cierta relación con sus posibilidades de éxito final en la asignatura, se integraron los datos obtenidos automática de Moodle con información proporcionada directamente por el profesorado responsable de las asignaturas. En este caso para cada estudiante se consideraron dos atributos adicionales, su calificación final en la escala oficial de la Universidad de Cantabria, y una variable booleana que indicaba si el estudiante superó o no con éxito la asignatura después de las evaluaciones finales. Es de destacar que en el caso de la asignatura en modalidad B-learning, la evaluación final se produce después de un examen final presencial y que cada uno de los 4 bloques temáticos también tenía asociado un examen parcial presencial.

Finalmente, otro de los objetivos que se planteaba en el trabajo era analizar la evolución a lo largo del tiempo del comportamiento de los estudiantes en los foros de la asignatura. Se trata básicamente de

comprobar si a medida que avanza el desarrollo de la asignatura los estudiantes se sentían más animados a participar activamente en los foros o si había periodos puntuales en la que la actividad global era mayor o menor. Para tales fines se optó por generar para cada estudiante un conjunto de variables que contabilizaban el número de mensajes leídos, de respuestas a mensajes y de nuevos debates abiertos en cada una de las quincenas en las que se dividía el curso. En el caso de la primera asignatura, teniendo en cuenta que tenía una duración de 4 meses más un mes adicional que correspondía al periodo de evaluaciones, se optó por generar 30 atributos adicionales, 3 correspondientes a cada una de las 10 quincenas. Estos atributos son los que se recogen en la tabla 3.4.

Atributo	Origen	Descripción
FL1,FL2,FL3,FL4,FL5 FL6,FL7,FL8,FL9,FL10	**FL**	Mensajes leídos
FR1,FR2,FR3,FR4,FR5 FR6,FR7,FR8,FR9,FR10	**FR**	Respuestas a otros mensajes
FN1,FN2,FN3,FN4,FN5 FN6,FN7,FN8,FN9,FN10	**FN**	Nuevos de debate

Tabla 3.4: Atributos de actividad quincenal en los foros.

3.4.4 .Transformación de Datos:

Transformar de los datos permite obtener otro espacio de atributos que represente la misma información de diferente manera. Este paso implica la transformación del tipo de algunos atributos, en

caso que fuera necesario, teniendo presente que convertir el tipo de un atributo a otro puede cambiar la semántica de dicho atributo. Este paso está muy ligado al algoritmo que procesará los datos para obtener conocimiento.

En este paso de transformación se realizó una normalización de los atributos numéricos para que todos se definan sobre una escala [0,1]. El valor mínimo de cada atributo se transforma en el 0, mientras que el valor máximo se convierte en 1. De esta manera se evita que los resultados de los algoritmos se vean condicionados por la diferencia de escala en los diferentes atributos. Por ejemplo, parece lógico que el número total de mensajes que un estudiante puede leer durante el desarrollo del curso sea bastante superior al número de mensajes que él mismo llega a escribir.

Tras ese proceso de normalización, la siguiente transformación fue la generación de un archivo en el formato ARFF. Este formato es el que utiliza el software Weka. Las figuras 3.3 y 3.4 muestran las capturas de pantalla de los archivos ARFF con los datos. En este caso únicamente se muestran las cabeceras de definición de los atributos.

Fig.3.3: Archivo en formato arff con los datos

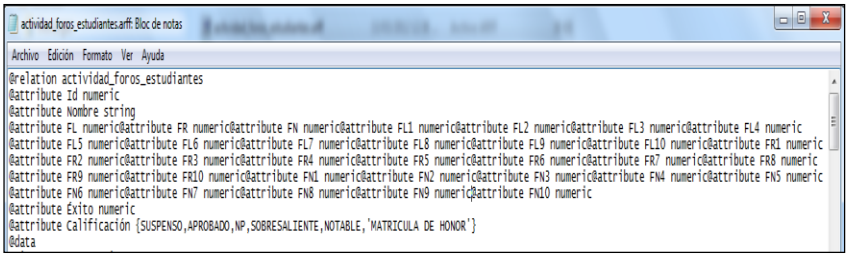


Fig.3.4: Archivo en formato arff con los datos quincenales

3.4.5 Reducción de Datos

En este paso se disminuye el tamaño de los datos, eliminando características redundantes. En este caso no fue necesario ningún proceso de reducción de datos.

IV RESULTADOS

4.1. MEDICIÓN DE NIVELES DE INTERACTIVIDAD UTILIZANDO EL PROCESO DE ANÁLISIS JERÁRQUICO (AHP)

Como se ha visto en las secciones previas, cada estudiante participando en los procesos de E-learning y B-learning tiene asociados una serie de atributos o indicadores relativos a su actividad en el curso virtual. Ahora bien, no todos los indicadores son igual de importantes a la hora de valorar el nivel de participación del estudiante en el curso. Por ejemplo, no es lo mismo participar en actividades que tienen un reflejo directo en la calificación (por ejemplo, entrega de tareas o realización de tests de evaluación) que participar en tareas en las que no ven ese reflejo directo (por ejemplo, participar en los foros de debate). Tampoco es lo mismo tener una participación pasiva en un foro (por ejemplo, leer mensajes) que tener una participación más activa (por ejemplo, responder a dudas de compañeros o plantear nuevos debates). En definitiva, a la hora de medir el grado de interactividad del estudiante con el curso virtual hay diferentes criterios a tener en cuenta. En este caso una metodología de decisión multicriterio como es la AHP resulta útil para la

obtención de un solo índice de interactividad a partir de los diferentes indicadores o criterios seleccionados. Además, también permite que incorporemos apreciaciones u opiniones subjetivas del profesor sobre los indicadores que pueden resultar más significativos cuando se trata de medir este grado de interactividad.

Como ya se comentó al presentar las metodologías a utilizar en esta tesis, AHP es una técnica de decisión multicriterio propuesta por Saaty (1980) para resolver problemas complejos de criterios múltiples. El proceso requiere que quien toma las decisiones proporcione evaluaciones subjetivas respecto a la importancia relativa de cada uno de los criterios y que después especifique su preferencia con respecto a cada uno de las alternativas de decisión para cada criterio. El resultado del AHP es una jerarquización con prioridades que muestran la preferencia global para cada una de las alternativas de decisión.

El AHP, mediante la construcción de un modelo jerárquico, permite de una manera eficiente y gráfica organizar la información respecto de un problema, descomponerla y analizarla por partes, visualizar los efectos de cambios en los niveles y sintetizar.

El AHP consta de cuatro etapas básicas:

1.- Modelización del problema.

2.-Evaluación y priorización de alternativas y criterios.

3.- Síntesis del problema.

4.- Análisis de sensibilidad

En las próximas secciones, tomando como referencia esas cuatro etapas básicas, se irán presentando los pasos realizados hasta la obtención de un conjunto de pesos que permitirán ponderar los indicadores de actividad de los estudiantes en el curso virtual. Para ello, además de mostrar el proceso de construcción algebraica, se utilizará el software Expert Choice, que implementa esta metodología de decisión.

4.1.1. Modelización del Problema

El modelo desarrollado tiene como objetivo lograr el objetivo básico de poder asignar a cada alumno un nivel de interactividad único, calculado a partir de los diferentes registros de actividad extraídos de la plataforma Moodle. Todos estos indicadores deben ser fácilmente cuantificables y no tener en cuenta los aspectos de rendimiento académico, simplemente la actividad.

Con el objetivo de comprobar la eficacia del modelo, se busca su aplicación práctica sobre las experiencias de B-learning llevadas a cabo sobre la plataforma Moodle de la Universidad de Cantabria.

4.1.1.1. Identificación de jerarquía de criterios

Tal como se ha comprobado en las secciones previas, haciendo uso de las funciones de administración de Moodle, el profesor puede obtener fácilmente la siguiente información sobre sus estudiantes:

- Cantidad de mensajes leídos en los foros de discusión o tablones de anuncios (RM).
- Número de respuestas dadas a mensajes de otros usuarios (MR).
- Número de nuevos temas de debate abiertos (ND).
- Número de días de actividad, con acceso efectivo a la plataforma (AD).
- Número de consultas realizadas a materiales curriculares (RV).
- Número de suscripciones a foros de discusión (S).
- Número de lecturas de mensajes e indicaciones de los profesores (RI).
- Número de veces en las que se ha participado en consultas o encuestas planteadas en el curso (PP).
- Número de ocasiones en las que se ha accedido a pruebas de evaluación en la plataforma (ET).
- Número de tareas entregadas a través de la plataforma (A).

Se decidió organizar jerárquicamente los 10 indicadores anteriores en tres categorías principales (ver Figura 4.1). La primera se relaciona con aspectos de interacción social, básicamente participación en foros. La segunda categoría está relacionada con la interacción con los materiales y recursos disponibles en el curso en línea. La última categoría de indicadores se relaciona con las tareas o actividades propuestas por el profesor a través de la plataforma.

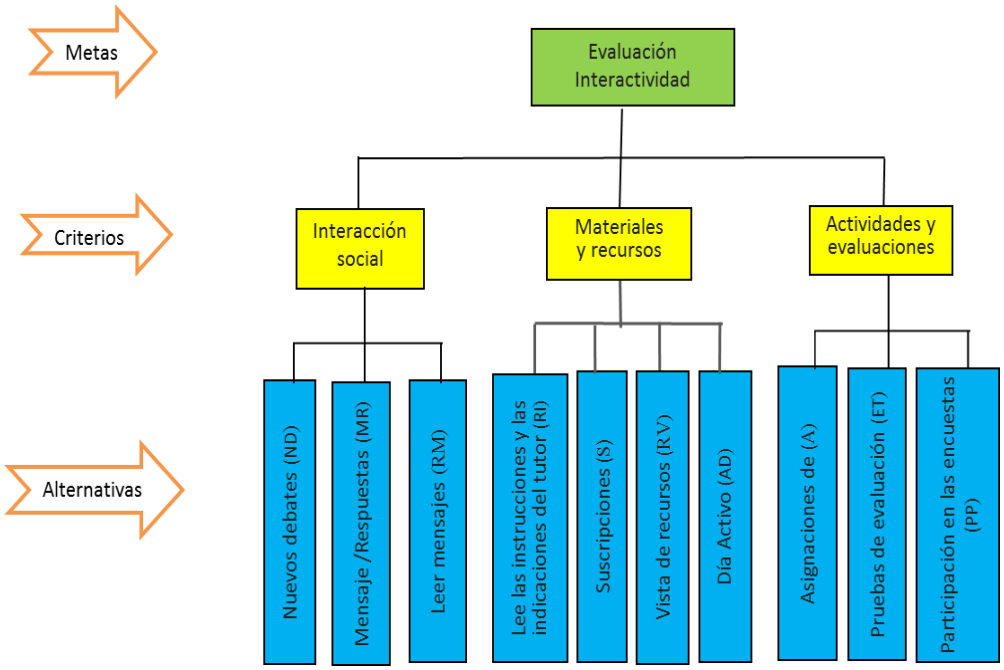


Fig.4.1: Jerarquía de criterios para elegir la interactividad del estudiante.

4.1.2. Evaluación y Priorización de Alternativas y Criterios.

Cuando se evalúa el nivel de interacción de un estudiante en un curso, cada profesor puede tener una percepción distinta en cuanto a qué criterios o indicadores son más significativos. Para ejemplo, ¿podríamos decir que un estudiante ha leído todos los mensajes enviados a los tablones de anuncios ha tenido un mayor nivel de interacción que otro que, a pesar de no haber leído muchos mensajes, ha planteado nuevos temas para los debates?

Por ello, se solicitó a los profesores responsables de las dos asignaturas analizadas que, de manera consensuada, procedieran a

realizar comparaciones dos a dos de la importancia de los criterios utilizando la escala de calificación de Saaty. Los resultados de la comparación de los criterios de nivel superior en la jerarquía son los que se muestran en la Tabla 4.1. Como puede verse, de acuerdo a los docentes, la interacción social (actividad en los foros) se considera como un indicador de mayor relevancia, aunque en un grado moderado (valores 2 y 3 en la escala Saaty) cuando se trata de medir la evaluación de la interacción de los estudiantes en los cursos on-line. En la matriz de la Tabla 4.1, por ejemplo, se está clasificado como 3 la importancia de la interacción social frente a la interacción con materiales y recursos del curso. De acuerdo al principio de reciprocidad, esa valoración implica que el valor de comparación entre materiales e interacción social sea de 1/3.

- **Comparación de a pares para los criterios**

	Interacción Social	**Materiales y Recursos**	**Actividades y evaluaciones**
Interacción Social	**1**	**3**	**2**
Materiales y Recursos	**1/3**	**1**	**1/2**
Actividades y evaluaciones	**½**	**2**	**1**

Tabla 4.1: Comparaciones por pares en el primer nivel jerárquico.

Una vez realizadas las comparaciones en el primer nivel, se realizó el mismo proceso de comparación de los diferentes subcriterios que se derivan de los criterios generales de la jerarquía. Las comparaciones obtenidas se muestran en las Tablas 4.2, 4.3 y 4.4. Por ejemplo, viendo estos datos podemos descubrir cómo al comparar la

importancia otorgada por los profesores a la puesta en marcha de la nueva debates por el estudiante (ND) con respecto a la simple lectura de los mensajes (RM), el primer criterio se considera mucho más importante (valor 7 en la escala).

	RM	MR	ND
RM	**1**	**1/5**	**1/7**
MR	**5**	**1**	**1/3**
ND	**7**	**3**	**1**

Tabla 4.2: Comparación por parejas de las alternativas según el criterio1

	AD	RV	S	RI
AD	**1**	**3**	**1**	**3**
RV	**1/3**	**1**	**1/2**	**1/2**
S	**1**	**2**	**1**	**3**
RI	**1/3**	**2**	**1/3**	**1**

Tabla 4.3. Comparación por parejas de las alternativas según el criterio 2

	PP	ET	A
PP	**1**	**3**	**3**
ET	**1/3**	**1**	**1**
A	**1/3**	**1**	**1**

Tabla 4.4: Comparación por parejas de las alternativas según el criterio 3

Una vez definidas esas tablas, y sus correspondientes matrices de valores, la metodología AHP pasa a realizar una valoración del nivel de consistencia lógica de las valoraciones dos a dos proporcionadas y a deducir unos pesos de ponderación a partir de las matrices. Ambas cosas pueden realizarse con el software Expert Choice.

4.1.3. Síntesis del problema

4.1.3.1. Razón de Consistencia

El AHP ofrece un método para medir el grado de consistencia entre las opiniones pareadas que proporciona el decisor. Si el grado de consistencia es aceptable, puede continuarse con el proceso de decisión. Si el grado de consistencia es inaceptable, quien toma las decisiones debe reconsiderar y posiblemente modificar sus juicios sobre las comparaciones pareadas antes de continuar con el análisis. Para medir la consistencia se considera la razón de consistencia (RC), considerándose aceptable un valor de RC menor o igual al 10%.

$\boldsymbol{RC} \leq 0.10$: Consistencia razonable
$\boldsymbol{RC} > 0.10$: Inconsistencia

El cálculo de la razón de consistencia para las comparaciones se realiza a partir de la siguiente expresión:

$$\mathrm{RC} = \frac{\mathrm{IC}}{\mathrm{IA}}$$

donde
IC= Índice de Consistencia
IA= Índice Aleatorio

El valor de IA se elige según la siguiente tabla

Nº de Elementos que se comparan	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Índice Aleatorio de Consistencia (IA)	0	0	0.58	0.89	1.11	1.24	1.32	1.40	1.45	1.49

Tabla. 4.5: Índice Aleatorio de Consistencia

El cálculo del IC está dado por

$$IC = \frac{\lambda_{max} - n}{n - 1}$$

Donde λ_{max} es el valor característico de mayor valor absoluto y n es la dimensión de la matriz.

El software Expert Choice realiza automáticamente esos cálculos al mismo tiempo que genera los pesos para cada criterio. Por ejemplo, la Figura 4.2 muestra como la razón de consistencia de las comparaciones de los criterios de primer nivel se sitúa en un valor de 0.0087, claramente por debajo de 0.1, por lo que las comparaciones deben ser aceptadas.

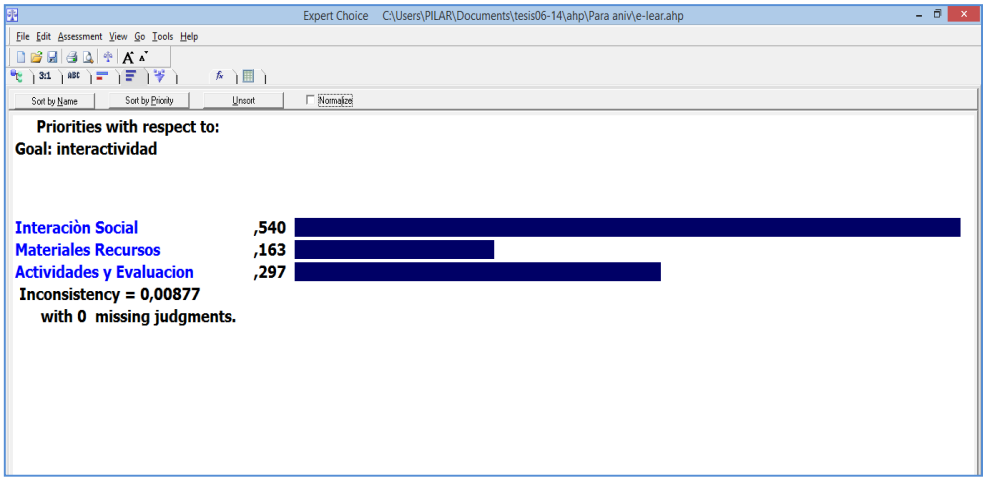


Fig. 4.2: Razón de Consistencia de las alternativas

De igual forma, las Figuras 4.3, 4.4 y 4.5 muestran razones de consistencia aceptables para las comparaciones del segundo nivel. Estas razones son de 0.06, 0.04 y 0.00.

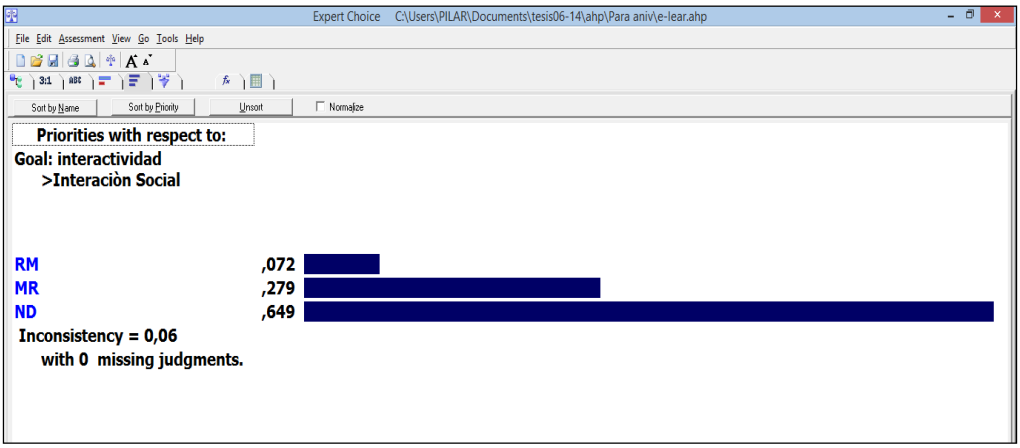


Fig. 4.3: Comparación por parejas de las alternativas según el criterio 1

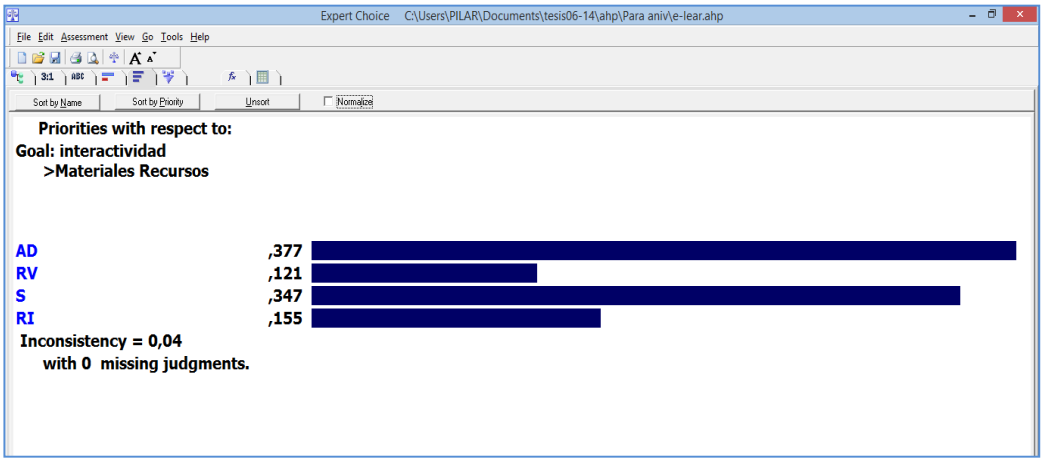


Fig. 4.4: Comparación por parejas de las alternativas según el criterio 2

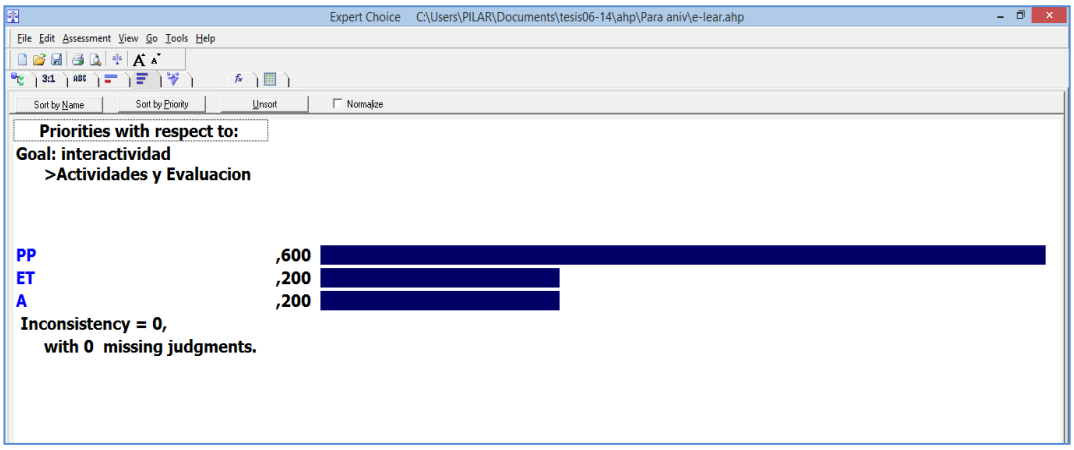


Fig. 4.5: Comparación por parejas de las alternativas según el criterio 3

4.1.3.2. Ponderación de los criterios

Además de calcular las razones de consistencia, el software Expert Choice deduce unos pesos de ponderación para cada criterio. Existen varios métodos de ponderación, el utilizado por Expert Choice consiste en un proceso iterativo de aproximación del autovector normalizado asociado al autovalor dominante. Ese autovalor es el que se utiliza también para medir el índice de consistencia. En las figuras 4.6, 4.7, 4.8 y 4.9 pueden verse los pesos asignados a cada criterio.

En especial conviene analizar los pesos que se generan a la hora de evaluar la interacción social de los estudiantes en los foros. Estos pesos serán utilizados más adelante para obtener un índice único de interactividad social para cada estudiante:

- Peso de la lectura de mensajes: 0.072
- Peso de las respuestas a otros mensajes: 0,279
- Pesos de la apertura de nuevos debates: 0,649

Como puede verse, más del 60% del peso total de esa interacción social recae sobre la actividad productora de contenidos en la forma de nuevos debates abiertos por los estudiantes.

Ahora bien, si se considera la ponderación global de todos los criterios con respecto a la medición de la interactividad total, se deberían multiplicar los correspondientes pesos de los subcriterios por los pesos de los criterios de nivel superior, obteniendo entonces los siguientes pesos que muestra la figura 4.7.

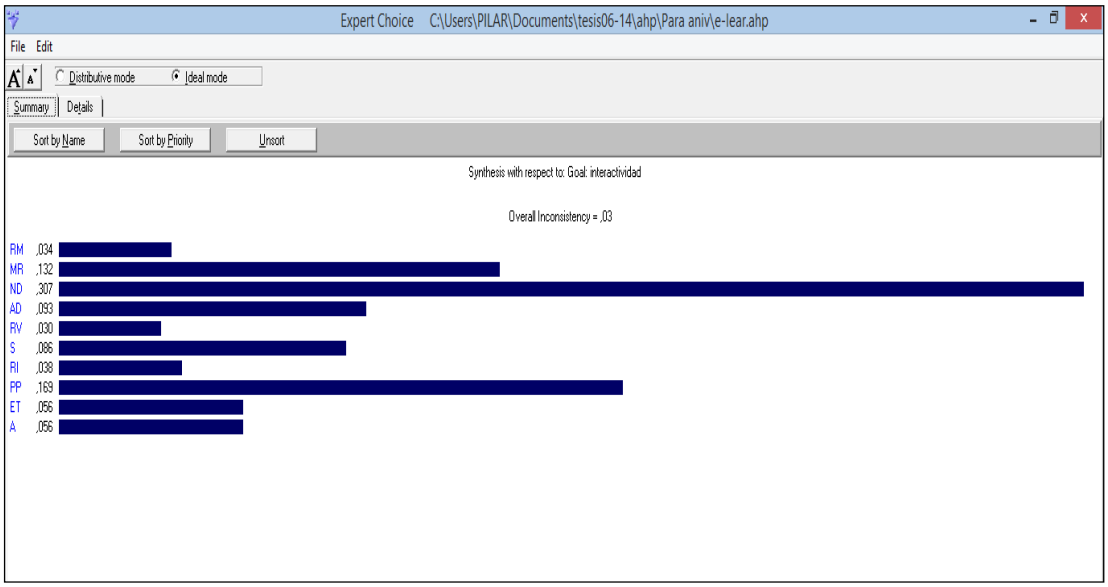


Fig. 4.7: Pesos relativos de cada criterio

Como se puede ver, los criterios que tienen mayor peso cuando se trata de evaluar el grado de interactividad son los que corresponden al inicio de nuevos debates (ND), la participación en encuestas propuestas por el profesor (PP) y respuestas a los mensajes y preguntas propuestas por los compañeros de clase (MR). Cabe destacar que éstos no tienen ninguna incidencia en la calificación final obtenida por el alumno en la asignatura. Puede resultar sorprendente que la participación en las discusiones de los foros reciba mayor peso que la entrega de tareas y evaluaciones. Sin embargo, hay que señalar que la realización de estas últimas actividades se encuentra ya incentivada por tener un impacto directo en la nota final.

Los pesos finales de todos los criterios se incluyen en la siguiente tabla:

Criterio	Peso
RM	0.034
MR	0.132
ND	0.307
AD	0.093
RV	0.030
S	0.086
RI	0.038
PP	0.169
ET	0.056
A	0.056

Tabla 4.6: Pesos de los Criterios

4.1.4. Análisis de Sensibilidad

Una de las ventajas importantes que presenta el AHP es la posibilidad de efectuar potentes análisis de sensibilidad, con el objetivo de localizar posibles modificaciones en la ordenación de alternativas ante variaciones en los pesos de los criterios. La utilización del Expert Choice en la resolución del problema de decisión permite realizar este tipo de análisis de una forma sencilla y fácil de interpretar gracias a la representación gráfica del problema observamos en la Fig.4.8.

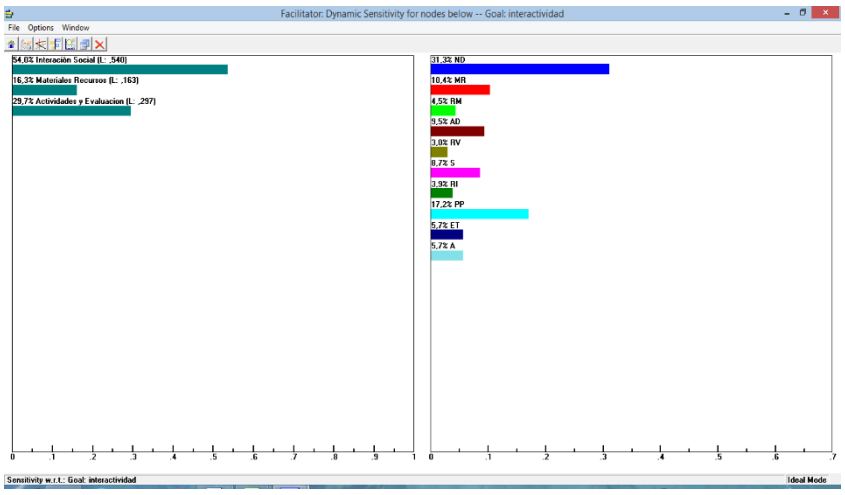


Fig. 4.8: Análisis de Sensibilidad

Los pesos calculados con AHP permitirían, por ejemplo, comparar la actividad global de los estudiantes. Si los valores de los correspondientes indicadores son ponderados por los correspondientes pesos y sumados todos ellos, se obtendría un único nivel de interactividad en el curso virtual. De esta manera se podría ordenar a los estudiantes según su interactividad. El análisis de sensibilidad en este caso permitiría experimentar cómo cambiaría esa ordenación si se modifican ligeramente los valores de esos pesos.

Además de aplicarse sobre los indicadores individuales de un estudiante, estos pesos pueden aplicarse sobre los valores promedios entre grupos de estudiantes. De esa manera se podrían ordenar los grupos de acuerdo a su nivel de interactividad con el curso. En la próxima sección se realiza precisamente esto sobre los grupos identificados automáticamente con un algoritmo de clustering.

4.2. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS SOBRE LOS INDICADORES DEL MODELO B-LEARNING

En esta sección se hace uso de un algoritmo de clustering para tratar de identificar automáticamente grupos de estudiantes que respondan a patrones similares en el uso de la plataforma de enseñanza virtual en la experiencia B-learning (asignatura de Teoría de la Optimización). Además se desea analizar si el comportamiento en el curso virtual y su nivel de participación guardan alguna relación con las probabilidades finales de éxito en la asignatura; es decir, saber si aquellos estudiantes que adoptan posiciones más activas en el curso virtual superan con mayor facilidad la asignatura. Como no se sabe a

priori cuántos grupos diferentes hay se opta por utilizar el algoritmo EM (Expectation Maximization). Este algoritmo de clústering encuentra la máxima verosimilitud (log-likelihood) mediante un método iterativo. Cada iteración se compone de dos pasos: la expectativa (expectation), donde se calculan las probabilidades de las instancias para conformar los grupos; la maximización, para calcular la distribución de los parámetros. De este modo, cuanto mayor sea el valor de log-likelihood, los grupos creados pueden considerarse mejores ya que log-likelihood es la métrica que permite comparar la calidad de los grupos, para cada ejecución del algoritmo EM.

Una vez identificados automáticamente los grupos, se utilizarán los pesos generados con AHP para ordenar los centroides de los grupos de acuerdo al nivel de interactividad global que tengan asociados. Eso permitirá identificar posteriormente los rasgos característicos de cada grupo. Para hacer esto se necesita asociar cada uno de los atributos que se utilizarán para la clasificación con los criterios que se utilizaron en el modelo AHP. La siguiente tabla 5.1 muestra esta asociación. Obsérvese que en alguno de los criterios no existen atributos seleccionados para los estudiantes que permitan obtener un valor para el mismo. En el caso de criterios asociados con más de un atributo, a la hora de la ponderación se considerará la media de los correspondientes atributos.

Código	Descripción	Atributos asociados	Peso
RM	Cantidad de mensajes leídos en los foros de discusión o tablones de anuncios	FGL, F1L, F2L, F3L, F4L	0.034
MR	Número de respuestas dadas a mensajes de otros usuarios	FGR, F1R, F2R, F3R, F4R	0.132
ND	Número de nuevos temas de debate abiertos	FGN, F1N, F2N, F3N, F4N	0.307
AD	Número de días de actividad, con acceso efectivo a la plataforma		0.093
RV	Número de consultas realizadas a materiales curriculares		0.030
S	Número de suscripciones a foros de discusión		0.086
RI	Número de lecturas de mensajes e indicaciones de los profesores	FNL	0.038
PP	Número de veces en las que se ha participado en consultas o encuestas planteadas en el curso		0.169
ET	Número de ocasiones en las que se ha accedido a pruebas de evaluación en la plataforma	T1, T2, T3, T4	0.056
A	Número de tareas entregadas a través de la plataforma	TR	0.056

Tabla.5.1: Criterios asociados con los atributos

A.- Actividades con el Algoritmo EM en el sistema B-learning

Como primer análisis se realizó un agrupamiento de estudiantes tomando como referencia 28 atributos para cada uno de ellos. En este caso a los atributos seleccionados e indicados previamente se le añadieron tres atributos relacionados con la participación en actividades presenciales en laboratorio de cómputo (LC1, LC2 y LC3) y los resultados de 4 pruebas de evaluación presenciales (TC1, TC2, TC3, TC4). Esos 7 atributos hacen referencia a rendimiento académico de los estudiantes pero no tanto a su nivel

de participación en el curso virtual de la asignatura. En definitiva, esos 7 atributos adicionales fueron utilizados para realizar el clustering de los estudiantes pero no para obtener un nivel de participación en el curso representativo de cada grupo.

En la Figura 5.1 aparece el resultado del proceso de clustering realizado con el algoritmo EM de Weka. Como puede apreciarse, el algoritmo identificó de forma automática 5 grupos de estudiantes. En esa figura se muestran los valores promedio de todos los atributos para los 5 grupos encontrados. Como indicador de calidad del algoritmo, se puede señalar que el valor de log-likelihood fue de 29.38892.

En cuanto al reparto de los estudiantes en los grupos, el primero de ellos está formado por 16 instancias (estudiantes), el segundo por 24 instancias, el tercero por 110 instancias, el cuarto por 7 instancias y el último, que además es el más números con el 50% de los estudiantes, por 154 instancias.

Cluster	0	1	2	3	4
N° de Estudiantes	16 (5%)	24 (8%)	110 (35%)	7 (2%)	154 (50%)
T1	0.7011	0.7434	0.7933	0.8499	0.6303
T2	0.7448	0.7518	0.8509	0.8994	0.6524
T3	0.6751	0.7198	0.723	0.8652	0.628
T4	0.6392	0.7182	0.8217	0.7957	0.671
LC1	0.5356	0.5365	0.4573	0.729	0.3595
LC2	0.4044	0.416	0.373	0.7651	0.2291
LC3	0.4577	0.3948	0.2689	0.5157	0.327
TR	0.9072	0.7092	0.4572	0.7144	0.6149
TC1	0.5956	0.5133	0.5169	0.775	0.3785
TC2	0.4044	0.416	0.373	0.7651	0.2291
TC3	0.3186	0.3138	0.3488	0.6303	0.2008
TC4	0.7229	0.658	0.5245	0.9222	0.4966
FNL	0.1984	0.1274	0.2274	0.3735	0.029
FGR	0.1665	0.0151	0.0397	0.0835	0.0019
FGN	0.2476	0.081	0.432	0.0523	0.0101
FGL	0.4573	0.1335	0.1745	0.4096	0.0253
F1L	0.1378	0.1808	0.2965	0.2247	0.0693
F2L	0.1588	0.2576	0.1546	0.5343	0.0895
F3L	0.118	0.0931	0.1079	0.3997	0.0365
F4L	0.2377	0.1154	0.1502	0.3622	0.0417
F1R	0.1238	0	0	0	0
F2R	0.1437	0.3523	0.2833	0.0034	0
F3R	0.2105	0.0046	0	0.4533	0
F4R	0.3057	0.008	0	0.3204	0
F1N	0.1238	0	0	0	0
F2N	0.2341	0	0	0	0
F3N	0.1548	0.0076	0	0	0
F4N	0.1574	0.0035	0	0	0

Fig.5.1. Actividades con el algoritmo EM

A la vista de los valores medios de los diferentes atributos en cada grupo, se pueden intuir una serie de características destacables de cada grupo. Además, tomando como referencia esos valores medios y la vinculación con los criterios del modelo AHP, se podría obtener un indicador único de nivel de actividad para cada grupo a partir de la siguiente fórmula:

$$NA = 0.034*Promedio(FGL,F1L,F2L,F3L,F4L) + 0.132*Promedio(FGR,F1R,F2R,F3R,F4R) + 0.307*Promedio(FGN,F1N,F2N,F3N,F4N) + 0.038*FNL + 0.056*Promedio(T1,T2,T3,T4) + 0.056*TR$$

Haciendo uso de esta fórmula sobre los centroides de cada grupo, la actividad virtual global de cada uno de ellos es la que se muestra a continuación:

Nivel de actividad virtual del Cluster 0	0,1860
Nivel de actividad virtual del Cluster 1	0,1066
Nivel de actividad virtual del Cluster 2	0,1199
Nivel de actividad virtual del Cluster 3	0,1410
Nivel de actividad virtual del Cluster 4	0,0741

Como puede observarse, hay dos grupos que destacan por tener un nivel de actividad virtual más alta que el resto: los clústeres 0 y 3. Si se comparan los valores de los indicadores de actividades presenciales en ambos grupos, se observa cómo los estudiantes del cluster 3 tienen valores más altos en los resultados de las actividades en el laboratorio de cómputo (LC1, LC2, LC3) y en las pruebas de evaluación presenciales de cada bloque temático (TC1, TC2, TC3, TC4).

El grupo más numeroso, el cluster 4, corresponde también con el que tiene un nivel de actividad virtual más bajo. Los resultados de actividades presenciales en este grupo son también más bajos que en el resto.

A la vista de todos los valores obtenidos se podría decir que las características de los diferentes grupos obtenidos serían las siguientes:

Grupo 0: Este grupo está formado por un 5% de los estudiantes. Son estudiantes con alta participación en las actividades propuestas a través del curso virtual. En especial son los que más participan en los foros respondiendo a consultas de los compañeros, pero también iniciando nuevos debates. Podríamos decir, por tanto, que responden a un perfil de estudiantes

"**Socializadores**". Sus resultados académicos son también buenos aunque no al nivel de otros grupos.

Grupos 1 y 2: Estos dos grupos responden a un perfil muy similar. En el primero se encuentran un 8% de los usuarios, mientras que en el segundo se sitúan el 35%. Ambos grupos se caracterizan por un nivel de participación medio en las actividades virtuales, por la poca actividad creativa en los foros, ya que apenas responden a mensajes o crean hilos de debate nuevos. Su rendimiento académico no se puede considerar malo pero su actividad virtual no es la óptima. Podríamos caracterizar a estos estudiantes como "**Medios o Grises**". Quizás los del grupo 1 obtienen unos resultados ligeramente mejores que los del grupo 2, pero en cambio, los del grupo 2 parecen estar más pendientes de los mensajes que se publican en los foros, con unos niveles de lectura ligeramente superiores. Estas ligeras diferencias pueden justificar que el algoritmo los haya separado en dos grupos diferentes, pero realmente el perfil de todos ellos es muy similar.

Grupo 3: En este grupo se encuentra un pequeño porcentaje de los estudiantes (un 2%). Son estudiantes con un buen nivel de participación en las actividades virtuales, especialmente en los foros. Además lo complementan con unos muy buenos resultados en las actividades presenciales, destacando claramente de los estudiantes del resto de grupos. Es por ello que podríamos considerar a este pequeño grupo de estudiantes como "**Líderes**".

Grupo 4: En este cluster se encuentra el grupo más numeroso de estudiantes, con un 50% de los inscritos en la asignatura. Parece que acceden a la plataforma únicamente en momentos puntuales, por ejemplo para realizar tests o para entregar tareas. Prácticamente no participan en los foros temáticos, ni leyendo los mensajes ni haciendo nuevas aportaciones. Los resultados académicos en promedio tampoco son destacados. En definitiva, se podría considerar a este grupo como estudiantes "**Desmotivados**", que no encuentran un suficiente incentivo para participar regularmente en el curso virtual o que no están lo suficientemente implicados en la asignatura.

La siguiente figura 5.2 muestra el reparto de los estudiantes en los 5 grupos generados por el algoritmo EM y los 4 perfiles de estudiantes que se podrían identificar a partir de las características de cada clúster.

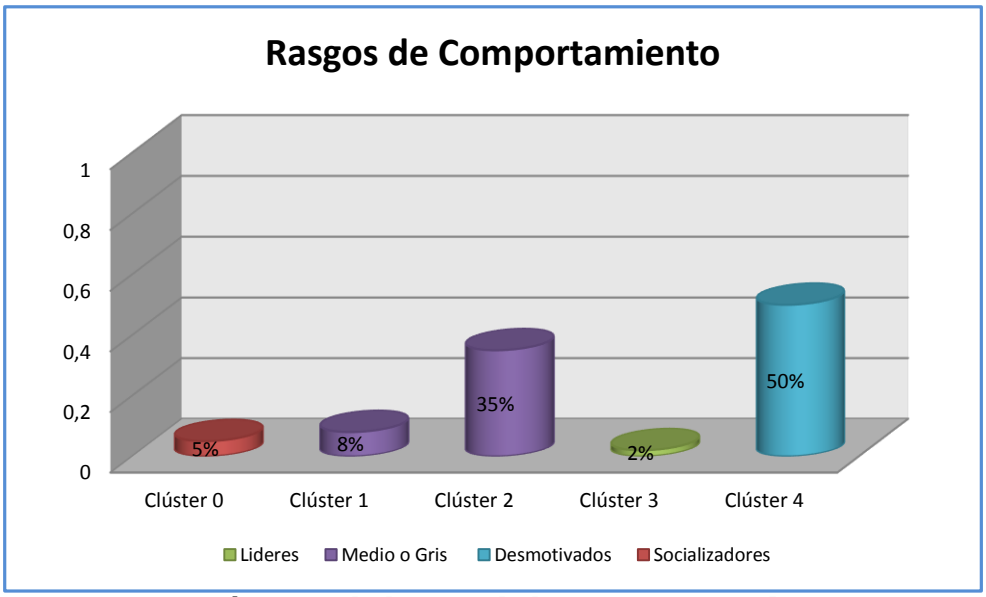


Fig.5.2: Rasgos según actividades con el algoritmo EM en el sistema B-learning

B.- Análisis de la actividad social de los estudiantes en los foros con el algoritmo SimpleK-Means

Como segundo análisis se plantea la identificación de patrones de comportamiento de los estudiantes en los foros de debate habilitados en la asignatura. En estos foros los estudiantes podían realizar básicamente tres tipos de actividades: leer mensajes, responder a mensajes de otros usuarios, o iniciar nuevos debates. Diversos estudios sobre interacción de usuarios en foros o en redes sociales sugieren la clasificación de cuatro tipos de usuarios:

1. Inactivos
2. Pasivos
3. Activos
4. Colaborativos

Globalmente, la clasificación anterior está planteada según el nivel de participación, de menor a mayor participación.

Con objeto de poder identificar esos perfiles entre los estudiantes de la asignatura, se optó por aplicar técnicas de clustering pero utilizando únicamente las variables de participación en los espacios de interacción social (foros) y no aquellas relacionadas con actividades planificadas y programadas por el profesor como son la entrega de tareas, actividades en laboratorios, realización de tests de evaluación. En concreto, los 16 atributos relacionados con la actividad en los foros son:

- Atributos de lectura de mensajes: FNL, FGL, F1L, F2L, F3L, F4L
- Atributos de respuestas a mensajes: FGR, F1R, F2R, F3R, F4R
- Atributos de generación de nuevos hilos: FGN, F1N, F2N, F3N, F4N.

Teniendo en cuenta que en este caso de antemano queremos identificar cuatro grupos de estudiantes, se recurrirá a la utilización de un algoritmo de clustering sencillo y rápido de ejecución que necesita como parámetro ese número de grupos: el algoritmo Simple k-Means.

Por otro lado, para ordenar los grupos de acuerdo a un nivel global de interactividad social, se utilizarán los pesos obtenidos con el modelo AHP para valorar los diferentes criterios. En concreto, los pesos son:

- Peso de la lectura de mensajes: 0.072
- Peso de las respuestas a otros mensajes: 0,279
- Pesos de la apertura de nuevos debates: 0,649

Con estos pesos, el indicador global de actividad social de cada estudiante o grupo se calcularía con la siguiente fórmula:

AS = 0.072*promedio(FNL,FGL,F1L,F2L,F3L,F4L) +
0.279*promedio(FGR,F1R,F2R,F3R,F4R) +
0.6449*promedio(FGN,F1N,F2N,F3N,F4N)

En la figura 5.3. se muestra el resultado usando el algoritmo SimpleK-Means divide los estudiantes en 4 grupos. En cuanto al reparto de los

estudiantes en los grupos, el primero de ellos está formado por 41 instancias (estudiantes), el segundo por 249 instancias, el tercero por 48 instancias, y el último, que además es el menor número con 4 instancias (estudiantes). Tal como se muestra a continuación:

Cluster N° de Estudiantes	0 (41) (12%)	1 249 (73%)	2 48 (14%)	3 4 (1%)
FNL	0.1266	0.0239	0.0919	0.5994
FGR	0.043	0.0095	0.0883	0.561
FGN	0.0461	0.0174	0.2685	0.36110
FGL	0.2713	0.0312	0.1026	0.0785
F1L	0.25	0.0409	0.1037	0.4927
F2L	0.1089	0.0206	0.0867	0.5429
F3L	0.164	0.019	0.1139	0.8027
F4L	0.0757	0.0079	0.0332	0.6027
F1R	0.0803	0.0123	0.0711	0.3529
F2R	0.0189	0.0048	0.0766	0.4919
F3R	0.021	0.0016	0.0458	0.4
F4R	0.0163	0.0033	0.0278	0.5
F1N	0.0366	0.0241	0.0938	0.125
F2N	0	0	0.3542	0.375
F3N	0	0	0.3229	0.125
F4N	0	0.004	0	0

Fig. 5.3: Foros con el algoritmo SimpleK-Means en el sistema B-learning

Igualmente haciendo uso de la fórmula anterior sobre los centroides de cada grupo, en foros global de cada uno de ellos es la que se muestra a continuación:

Nivel de actividad virtual del Clúster 0	0,0326
Nivel de actividad virtual del Clúster 1	0,0093
Nivel de actividad virtual del Clúster 2	0,1577
Nivel de actividad virtual del Clúster 3	0,2933

Como puede observarse, el grupo que con el nivel de actividad social que destacan es el clúster 3 más alta que el resto; se observa cómo los estudiantes en este grupo tienen mayor participación en los mensajes leídos (FNL,F1L,F2L,F3L,F4L), y en las respuestas a otros mensajes (FGR,F1R,F2R,F3R,F4R) .

El segundo grupo con actividad social más alta es el correspondiente al Cluster 2, pero con buenos niveles de lectura de mensajes y participación en debate nuevo; sin embargo con poca participación de colaboración a través de respuestas a compañeros.

El grupo más numeroso, el cluster 1, corresponde también con el que tiene un nivel de actividad social más bajo. Los resultados de la participación en los foros en este grupo son también más bajos que en el resto.

A la vista de todos los valores obtenidos se podría decir que las características de los diferentes grupos obtenidos serían las siguientes:

Grupo 0: Este grupo está formado por un 12% de los estudiantes con poca participación en los foros leídos, respondidos a consultas de los compañeros, pero nula en nuevos debates. Podríamos decir, por tanto, que responden a un perfil de estudiantes "**Pasivos**".

Grupo 1: En este cluster se encuentra el grupo más numeroso de estudiantes, con un 73% de los inscritos en la asignatura, participación medio en las actividades sociales, por la poca actividad creativa en los foros, ya que

apenas responden a mensajes o crean hilos de debate nuevos. Podríamos caracterizar a estos estudiantes como "**Inactivos**".

Grupo 2: Este grupo está constituido por un 14% estudiante con actividad social alta, Es por ello que podríamos considerar a este pequeño grupo de estudiantes como "**Activos**" pero con buenos niveles de lectura de mensajes y de colaboración a través de respuestas a compañeros; sin embargo no han planteado ningún debate nuevo.

Grupo 3: En este grupo se encuentra un pequeño porcentaje de los estudiantes (un 1%) con el nivel de actividad social más altos, con buenos niveles de lectura de los mensajes de los foros y de respuesta a mensajes de compañeros, pero sobre todo con alto nivel en el planteamiento de nuevos hilos de debate. Este grupo corresponde, por tanto al perfil de estudiantes **Colaborativos**.

La siguiente figura 5.5 muestra el reparto de los estudiantes en los 4 grupos generados por el algoritmo Simplek-means y los 4 perfiles de estudiantes que se podrían identificar a partir de las características de cada clúster.

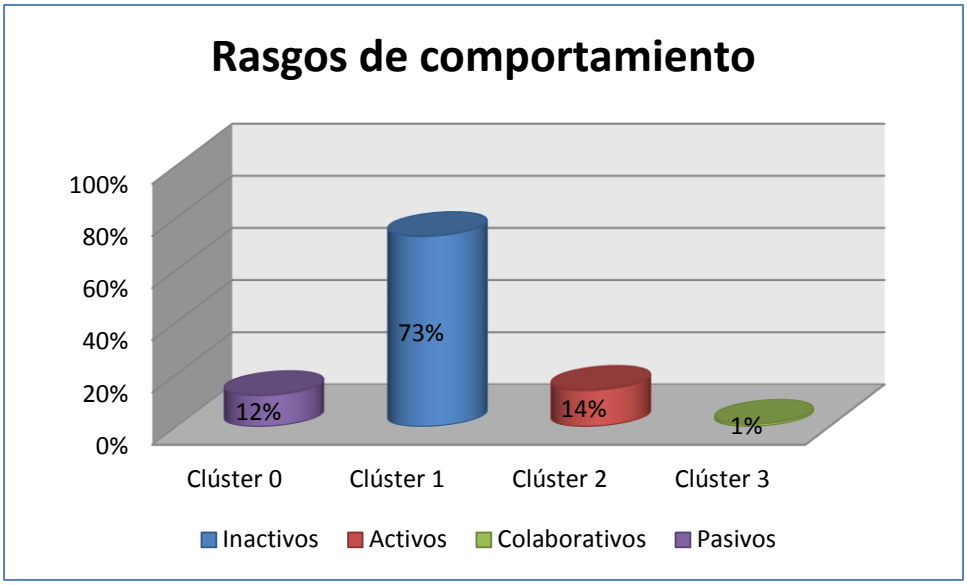


Fig.5.4: Rasgos de comportamiento de estudiantes según su nivel de participación en los foros.

C.- Análisis temporal de la actividad social de los usuarios: clustering con el algoritmo SimpleK-Means sobre la actividad quincenal.

En esta sección se tratará de analizar la evolución de los rasgos de comportamiento de los estudiantes con respecto a la participación en los foros a medida que avanza el desarrollo de la asignatura. Para ello se utilizarán indicadores de actividad social por quincenas. Para cada una de las quincenas en las que se divide el curso se cuenta con tres atributos de cada estudiante:

1. Número de mensajes leídos en la quincena.
2. Número de respuestas a otros mensajes en la quincena.
3. Número de nuevos mensajes creados en la quincena.

A partir de esos tres atributos se puede generar un nivel de actividad social de la quincena mediante la expresión resultante de ponderar los atributos con los pesos obtenidos en el modelo AHP, es decir:

$$AS(k) = 0.072*FL_k + 0.279*FR_k + 0.6449*FN_k$$

siendo k el número de quincena.

De nuevo, al buscarse la clasificación en las cuatro categorías de usuarios anteriores (inactivos, pasivos, activos y colaborativos) se usa el algoritmo Simple k-Means indicando que construya 4 grupos.

En las próximas secciones se presentan los resultados obtenidos en cada una de las quincenas.

1.- Primera Quincena

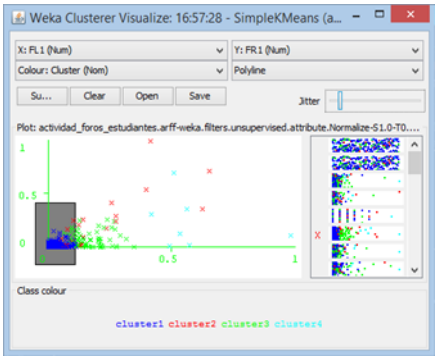
En la primera quincena se observa un agrupamiento un tanto especial ya que el 71% de los estudiantes son asignados a un Cluster con valores muy bajos en los atributos. Se puede entender al ser las dos primeras semanas de la asignatura con bastantes estudiantes que aún no toman actitudes realmente de trabajo en la asignatura, algunos otros que están aún regularizando su situación administrativa en el inicio del semestre y otros que simplemente se toman unos días más relajados tras salir del periodo de exámenes con el que se cerraba el semestre precedente. En concreto la composición de los 4 grupos creados es la que se recoge en la siguiente tabla:

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                 (342)      (243)      (23)       (69)       (7)
=============================================================================
FL1            0.0801     0.0277     0.2007      0.173     0.5865
FR1            0.0327     0.0034     0.2097      0.058     0.2185
FN1            0.0365          0     0.5435          0          0


Clustered Instances

0      243 ( 71%)
1       23 (  7%)
2       69 ( 20%)
3        7 (  2%)
```

Participación en los foros durante la 1era quincena



Clustering visual

Tomando como referencia los centroides de cada grupo, puede calcularse la actividad social promedio durante la primera semana *AS(1)* para los 4 grupos:

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL1	0,0277	0,2007	0,173	0,5865
FR1	0,0034	0,2097	0,058	0,2185
FN1	0	0,5435	0	0
AS(1)	0,0029	0,4235	0,0286	0,1032
Perfil	Inactivos	Colaborativos	Pasivos	Activos

Actividad social y perfil en la 1era quincena

El grupo con el nivel de actividad social más alto es el Cluster 1, formado por 23 estudiantes (un 7% del total) con buenos niveles de lectura

de los mensajes de los foros y de respuesta a mensajes de compañeros, pero sobre todo con alto nivel en el planteamiento de nuevos hilos de debate. Este grupo 1 corresponde, por tanto al perfil de estudiantes **Colaborativos**. El segundo grupo con actividad social más alta es el correspondiente al Cluster 3, este es por tanto el grupo de estudiantes que responden a un perfil **Activo**. Se trata de un grupo pequeño, con solo 7 estudiantes (un 2% del total) pero con buenos niveles de lectura de mensajes y de colaboración a través de respuestas a compañeros; sin embargo no han planteado ningún debate nuevo. Siguiendo con el orden de mayor a menor nivel de actividad social, el siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 2, con 69 estudiantes (un 20% del total) con cierta actividad de lectura de mensajes pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros. Finalmente, como ya se comentó, el grupo con nivel de actividad más bajo es el Cluster 0, con 243 estudiantes (71% del total) que prácticamente no accedieron a los foros. Estos estudiantes son los que responden a un perfil **Inactivo**.

2.- Segunda Quincena

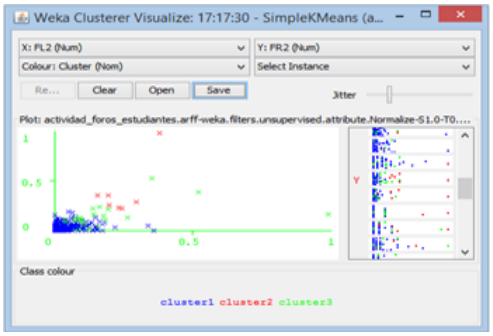
El análisis realizado en la primera quincena se replica exactamente igual en las restantes quincenas del semestre en el que se impartió la asignatura. A continuación se recogen los resultados correspondientes a la segunda quincena:

```
                           Cluster#
Attribute    Full Data          0          1          2
                 (342)      (304)        (7)       (31)
=============================================================
FL2               0.0466     0.0315     0.2422     0.1502
FR2               0.0223     0.0047     0.3963     0.1103
FN2               0.0541          0          0     0.5968


Clustered Instances

0       304 ( 89%)
1         7 (  2%)
2        31 (  9%)
```

Participación en los foros durante la 2ª quincena



Clustering visual

La actividad social y perfiles de los grupos es la que se muestra a continuación:

Atributos	Cluster 0	Cluster 1	Cluster 2
FL2	0,0315	0,2422	0,1502
FR2	0,0047	0,3963	0,1103
FN2	0	0	0,5968
AS(2)	0,0036	0,1280	0,4265
Perfil	Inactivos	Activos	Colaborativos

Actividad social y perfil en la 2da quincena

El grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 2, que agrupa a 31 estudiantes

(9%). Estos son los únicos estudiantes que plantearon nuevos mensajes en los foros. En un segundo nivel de actividad se considera estudiantes **Activos** se encuentra el Cluster 1, con 7 estudiantes (2%) con buenos niveles de lectura y de respuestas a mensajes de compañeros. De nuevo, el grupo con menor índice de actividad global es el Cluster 0 con 304 estudiantes **Inactivos** durante la quincena. Son estudiantes con un nivel de acceso a los foros muy pequeño.

3.- Tercera Quincena

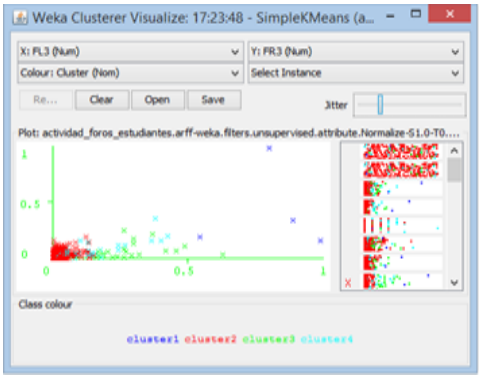
En tercera quincena los resultados obtenidos son los que se muestran a continuación:

		Cluster#			
Attribute	Full Data	0	1	2	3
	(342)	(4)	(290)	(21)	(27)
FL3	0.0588	0.8027	0.0228	0.2963	0.1514
FR3	0.0148	0.4	0.0029	0.0452	0.062
FN3	0.0468	0.125	0	0	0.5741

Clustered Instances

0	4	(1%)
1	290	(85%)
2	21	(6%)
3	27	(8%)

Participación en los foros durante la 3ª quincena



Clustering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL3	0,8027	0,0228	0,2963	0,1514
FR3	0,4	0,0029	0,0452	0,062
FN3	0,125	0	0	0,5741
AS(3)	0,2500	0,0025	0,0339	0,3984
Perfil	Activos	Inactivos	Pasivos	Colaborativos

Actividad social y perfil en la 3era quincena

El grupo de más alto nivel de actividad social estudiantes **Colaborativos** es el correspondiente al Cluster 3, que agrupa a 27 estudiantes (8%). Estos son estudiantes que plantearon nuevos mensajes en los foros, así como los foros leídos y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 0, con 4 estudiantes (1%) con buenos niveles de lectura, respuestas a mensajes de compañeros y creación a nuevos mensajes. Siguiendo con el orden de mayor a menor nivel de actividad social, el siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 2, con 21 estudiantes (un 6% del total) con cierta actividad de lectura de mensajes pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros. De nuevo, el grupo con menor índice de actividad global es

el Cluster 1 con 390 estudiantes (85%) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros muy pequeño.

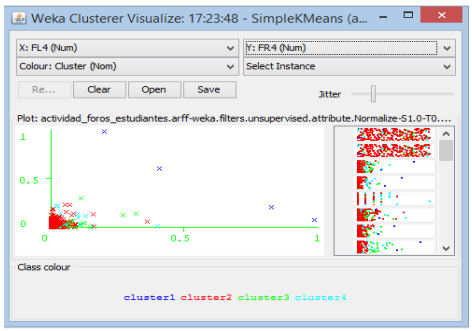
4.- Cuarta Quincena

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                 (342)        (5)       (15)      (280)       (42)
=============================================================================
FL4             0.0266     0.5119     0.0851     0.0046     0.0941
FR4             0.0141     0.4333     0.1778          0          0
FN4             0.0029        0.2          0          0          0


Clustered Instances

0         5 (  1%)
1        15 (  4%)
2       280 ( 82%)
3        42 ( 12%)
```

Foros en la 4ta quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL4	0,5119	0,0851	0,0046	0,0941
FR4	0,4333	0,1778	0	0
FN4	0,2	0	0	0
AS(2)	0,2867	0,0557	0,0003	0,0068
Perfil	Colaborativos	Activos	Inactivos	Pasivos

Actividad social y perfil en la 4ta quincena

En la 4ta quincena el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 0, que agrupa a 5 estudiantes (1%), estos son estudiantes que plantearon nuevos mensajes, y con buenos niveles en los foros, leídos y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 1, con 15 estudiantes (4%) con buenos niveles de respuestas a mensajes de compañeros y lectura de mensajes pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 3, con 42 estudiantes (un 12% del total) con cierta actividad de lectura de mensajes pero directamente nula actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros. De nuevo, el grupo con menor índice de actividad global es el Cluster 2 con 280 estudiantes (82 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño pero nula en los mensajes nuevos y respondidos.

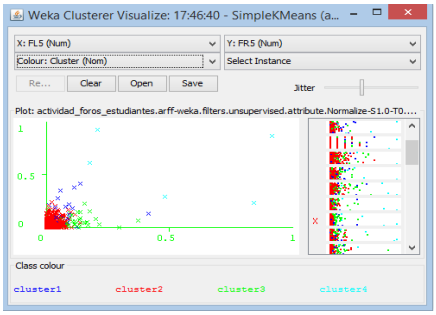
5.- Quinta Quincena

		Cluster#			
Attribute	Full Data	0	1	2	3
	(342)	(18)	(283)	(36)	(5)
FL5	0.0312	0.0938	0.0091	0.1048	0.5299
FR5	0.02	0.1574	0.0006	0.0301	0.55
FN5	0.0117	0.125	0	0	0.35

Clustered Instances

0	18	(5%)
1	283	(83%)
2	36	(11%)
3	5	(1%)

Foros en la 5ta quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL5	0,0938	0,0091	0,1048	0,5299
FR5	0,1574	0,0006	0,301	0,55
FN5	0,125	0	0	0,35
AS(5)	0,1313	0,0008	0,0915	0,4173
Perfil	Activos	Inactivos	Pasivos	Colaborativos

Actividad social y perfil en la 5ta quincena

En esta 5ta quincena el grupo de más alto nivel de actividad social son estudiantes Colaborativos es el correspondiente al Cluster 3, que agrupa a 5 estudiantes (1%), estos son los estudiantes con buen niveles de participación a los mensajes leídos, respuestas a mensajes de compañeros que plantearon y nuevos mensajes en los foros, leídos y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 0, con 18 estudiantes (5%) con alta participación a los mensajes leídos, respuestas a mensajes de compañeros y nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 1, con 283 estudiantes (un 83% del total) con cierta actividad de lectura de mensajes, respuestas a mensajes de compañeros, pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes. El grupo con menor índice de

actividad global es el Cluster 2 con 36 estudiantes (11 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño pero nula en los mensajes nuevos y respondidos.

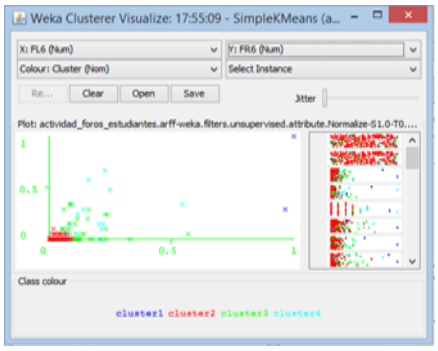
6.- Sexta Quincena

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                 (342)        (2)      (285)       (38)       (17)
=============================================================================
FL6             0.0407      0.983     0.0127     0.1443     0.1668
FR6             0.0259     0.6429     0.0022     0.0865     0.2157
FN6             0.0497          0          0          0          1


Clustered Instances

0        2 (  1%)
1      285 ( 83%)
2       38 ( 11%)
3       17 (  5%)
```

Foros en la 6ta quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL6	0,983	0,0127	0,1443	0,1668
FR6	0,6429	0,0022	0,0865	0,2157
FN6	0	0	0	1
AS(6)	0,2501	0,0015	0,0345	0,7171
Perfil	Activos	Inactivos	Pasivos	Colaborativo

Actividad social y perfil en la 6ta quincena

En esta 6ta quincena el grupo de más alto nivel de actividad social son estudiantes Colaborativos es el correspondiente al Cluster 3, que agrupa a 17 estudiantes (5%), estos son los estudiantes que plantearon nuevos mensajes en los foros, leídos y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 0, con 2 estudiantes (1%) con buenos niveles de mensajes leídos, respuestas a mensajes de compañeros pero nulo en los nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Clúster 2, con 38 estudiantes (un 11% del total) con cierta actividad de lectura de mensajes, respuestas a mensajes de compañeros, pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Clúster 1 con 285 estudiantes (83 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño y respondidos, pero nula en los mensajes nuevos.

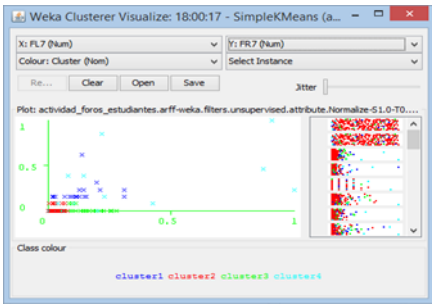
7.- Séptima Quincena

		Cluster#			
Attribute	Full Data	0	1	2	3
	(342)	(20)	(283)	(23)	(16)
FL7	0.04	0.132	0.0103	0.1565	0.2836
FR7	0.0261	0.2	0.0024	0.01	0.25
FN7	0.0468	0	0	0	1

Clustered Instances

0	20	(6%)
1	283	(83%)
2	23	(7%)
3	16	(5%)

Foros en la 7ma quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL7	0,132	0,0103	0,1565	0,2836
FR7	0,2	0,0024	0,01	0,25
FN7	0	0	0	1
AS(7)	0,0653	0,0014	0,0141	0,7351
Perfil	Activos	Inactivos	Pasivos	Colaborativos

Actividad social y perfil en la 7ma quincena

En esta 7ma quincena el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 3, que agrupa a 16 estudiantes (5%), estos son estudiantes que participaron en los mensajes leídos y respondidos. En un segundo nivel de actividad se considera estudiantes **Activos** se encuentra el Cluster 0, con 20 estudiantes (6%) con buenos niveles de mensajes leídos, respuestas a mensajes de compañeros pero nulo en los nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 2, con 23 estudiantes (un 7% del total) con cierta actividad de lectura de mensajes, respuestas a mensajes de compañeros, pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Cluster 1 con 283 estudiantes (83 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes

con un nivel de acceso a los foros leídos muy pequeño y respondidos, pero nula en los mensajes nuevos.

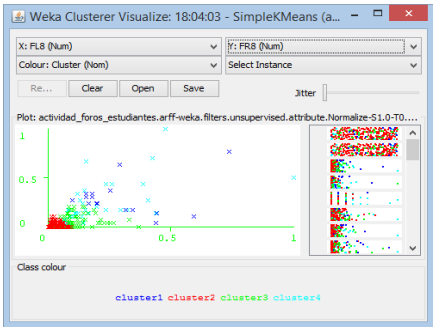
8.- Octava Quincena

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                 (342)       (17)      (237)       (56)       (32)
=============================================================================
FL8             0.0696     0.2887     0.0182     0.1353     0.2185
FR8             0.0561     0.2843     0.0094     0.0738       0.25
FN8               0.04          0          0          0     0.4271


Clustered Instances

0       17 (  5%)
1      237 ( 69%)
2       56 ( 16%)
3       32 (  9%)
```

Foros en la 8va quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL8	0,2887	0,0182	0,1353	0,2185
FR8	0,2843	0,0094	0,0738	0,25
FN8	0	0	0	0,4271
AS(8)	0,1001	0,0039	0,0303	0,3609
Perfil	Activos	Inactivos	Pasivos	Colaborativos

Actividad social y perfil en la 8va quincena

En esta 8va quincena el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 3, que agrupa a 32

estudiantes (9%), estos son estudiantes que plantearon nuevos mensajes en los foros, leídos y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 0, con 17 estudiantes (5%) con buenos niveles de mensajes leídos, respuestas a mensajes de compañeros pero nulo en los nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 2, con 56 estudiantes (un 16% del total) con cierta actividad de lectura de mensajes, respuestas a mensajes de compañeros, pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Cluster 1 con 237 estudiantes (69 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño y respondidos, pero nula en los mensajes nuevos.

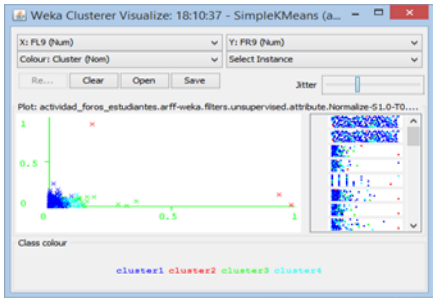
9.- Novena Quincena

		Cluster#			
Attribute	Full Data	0	1	2	3
	(342)	(263)	(3)	(14)	(62)
FL9	0.0229	0.0004	0.7265	0.1731	0.0502
FR9	0.0029	0	0.3333	0	0
FN9	0	0	0	0	0

Clustered Instances

0	263	(77%)
1	3	(1%)
2	14	(4%)
3	62	(18%)

Foros en la 9na quincena



Clústering visual

tributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL9	0,0004	0,7265	0,1731	0,0502
FR9	0	0,3333	0	0
FN9	0	0	0	0
AS(9)	0,0000	0,1453	0,0125	0,0036
Perfil	Inactivos	Colaborativos	Activos	Pasivos

Actividad social y perfil en la 9na quincena

En esta 9na quincena el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 1, que agrupa a 3 estudiantes (1%), se considera a este grupo estudiantes con un nivel de acceso alta en los mensajes leídos y respondidos, pero nulos en los nuevos mensajes en los foros. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 2, con 14 estudiantes (4%) con buenos niveles de mensajes leídos, pero nulo a mensajes respondidos y los nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 3, con 62 estudiantes (un 18% del total) con buenos niveles de mensajes leídos, pero nulo a mensajes respondidos, y los nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Cluster 1 con 263 estudiantes (67 %) considerándose estudiantes **Inactivos** durante esta

quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño, pero nula en los mensajes nuevos y respondidos.

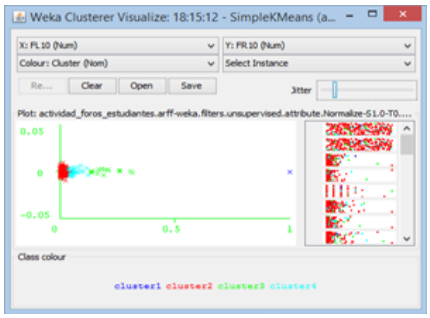
10.- Decima Quincena

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                 (342)        (1)      (293)      (14)       (34)
=============================================================================
FL10               0.0176       1     0.0006     0.2009     0.06
FR10                    0       0          0          0        0
FN10                    0       0          0          0        0


Clustered Instances

0         1 (  0%)
1       293 ( 86%)
2        14 (  4%)
3        34 ( 10%)
```

Foros en la 10ma quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL10	1	0,0006	0,2009	0,06
FR10	0	0	0	0
FN10	0	0	0	0
AS(10)	0,0720	0,0000	0,0145	0,0043
Perfil	Colaborativos	Inactivos	Activos	Pasivos

Actividad social y perfil en la 10ma quincena

En esta 10ma quincena el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 0, que agrupa a 1 estudiantes (0%), con buenos niveles de participación en los mensajes leídos, pero nulos en los nuevos mensajes en los foros y respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Clúster 2, con 14 estudiantes (4%) con buenos niveles de mensajes leídos, pero nulo a mensajes respondidos y los nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**. En este caso se trata del Cluster 3, con 34 estudiantes (un 10% del total) con buenos niveles de mensajes leídos, pero nulo a mensajes respondidos, y los nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Cluster 1 con 293 estudiantes (86%) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño, pero nula en los mensajes nuevos y respondidos.

D.- Resumen de los foros quincenal del Algoritmo Simplek-Means en el Sistema B-learning

En la tabla 5.2 podemos ver la evolución de los rasgos de comportamiento de los estudiantes en su participación en los foros del curso virtual a los largo de las 4 quincenas.

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
1era quincena	243	23	69	7
	71%	7%	20%	2%
	Inactivos	Colaborativos	Pasivos	Activos
2da quincena	304	7	31	
	89%	2%	9%	
	Inactivos	Activos	Colaborativos	
3ra quincena	4	290	21	27
	1%	85%	6%	8%
	Activos	Inactivos	Pasivos	Colaborativos
4ta quincena	5	15	250	42
	1%	5%	82%	12%
	Colaborativos	Activos	Inactivos	Pasivos
5ta quincena	18	283	36	5
	5%	83%	11%	1%
	Activos	Inactivos	Pasivos	Colaborativos
6ta quincena	2	285	38	17
	1%	83%	11%	5%
	Activos	Inactivos	Pasivos	Colaborativos
7ma quincena	20	283	23	16
	6%	83%	7%	5%
	Activos	Inactivos	Pasivos	Colaborativos
8va quincena	17	237	56	32
	5%	69%	16%	9%
	Activos	Inactivos	Pasivos	Colaborativos
9na quincena	263	3	14	62
	77%	1%	4%	18%
	Inactivos	Colaborativos	Activos	Pasivos
10ma quincena	1	293	14	34
	0%	86%	4%	10%
	Colaborativos	Inactivos	Activos	Pasivos

Tabla Nº 5.2: Foro quincenal con el algoritmo Simplesk-Means en el sistema B-learning

De la misma manera, la figura 5.5 muestra también esa evolución en la actividad social en los foros del curso virtual durante el desarrollo de la asignatura.

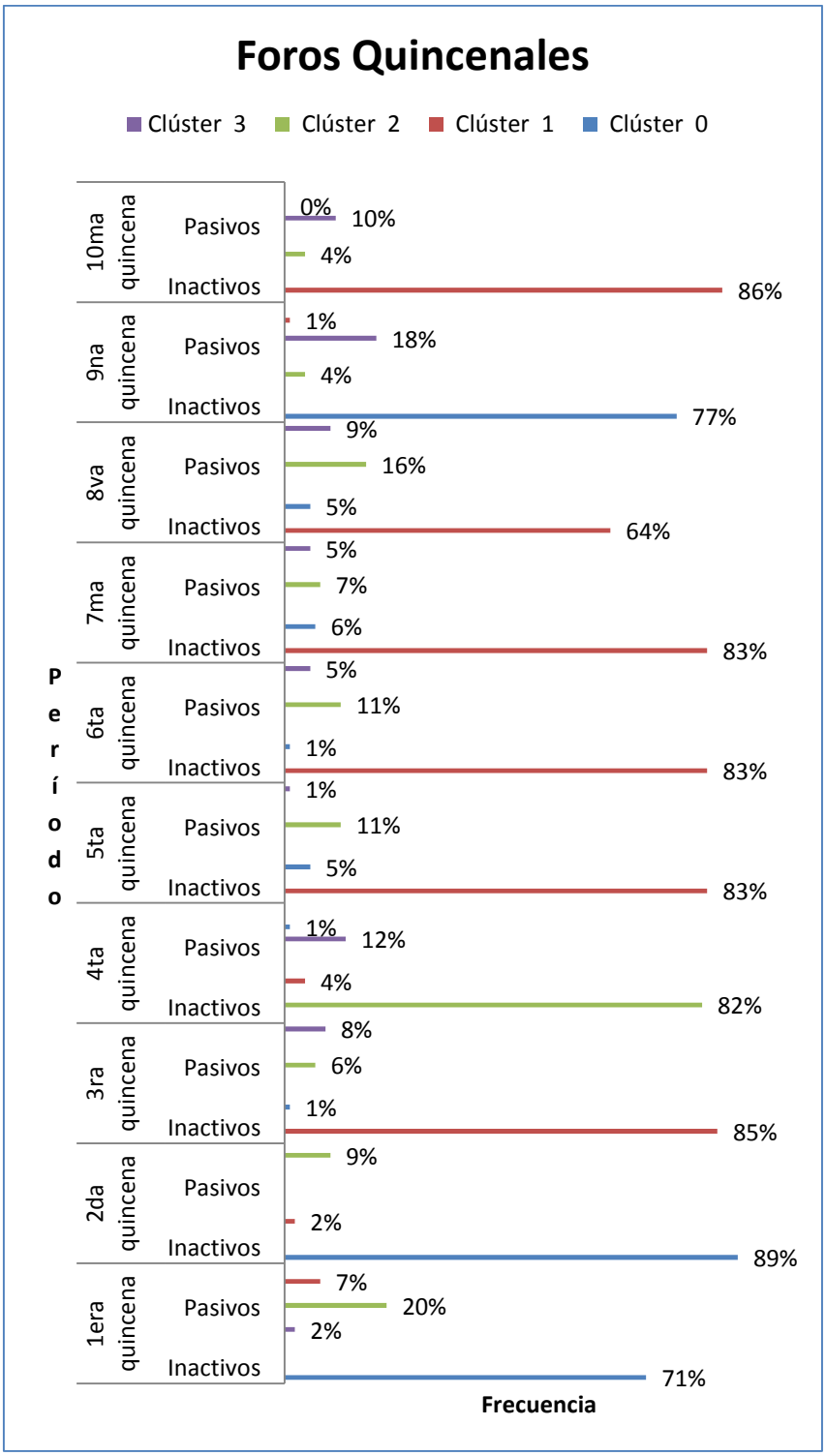


Fig. 5.5: Foro quincenal con el algoritmo Simplesk-Means en el sistema B-learning

En la Fig. 5.5 Se concluye durante las 10 quincenas, los estudiantes con participación en los foros durante el desarrollo del curso B-learning, con el Algoritmo Simplek-Means, en la primera quincena se observa un agrupamiento un tanto especial ya que el 71% de los estudiantes son asignados a un Cluster con valores muy bajos en los atributos. Se puede entender al ser la primera semana de la asignatura con bastantes estudiantes que aún no toman actitudes realmente de trabajo en la asignatura, continuando con las 2da , 3era, 4ta, quincenas se muestran estudiantes ***Inactivos*** por el poco interés en la participación en los foros. Parece que acceden a la plataforma únicamente en momentos puntuales, por ejemplo para realizar tests o para entregar tareas. Prácticamente no participan en los foros temáticos, ni leyendo los mensajes ni haciendo nuevas aportaciones.

Igualmente se observa en la 5ta quincena estudiantes **Inactivos** por la participación nula en todos los foros, pues se justifica debido a que los estudiantes se encuentra en un periodo de vacaciones de 10 días, en cuanto a la 6ta está reiniciando las clases tampoco acceden a la plataforma considerando también estudiantes **Inactivos.**

Ya en 7ma ,8va y 9na quincena se observa más participación de estudiantes **Pasivos** pues esto indica que por el motivo de realizar tests o para entregar tareas, sin embargo en las actividades sociales, su acceso a los foros temáticos tiene buenos niveles en los mensajes leídos, pero nulo a mensajes respondidos y los nuevos mensajes. Finalmente en la 10ma se observa el mayor porcentaje

(89%) de los estudiantes **Inactivos** con valores muy bajos en los atributos, podíamos decir que se debe que se está finalizando el curso B-learning, en forma general que en esta modalidad el desarrollo del curso es mixto es decir virtual y presencial vemos poca participación en las actividades sociales.

4.3. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS SOBRE LOS INDICADORES DEL MODELO E-LEARNING

Con objeto de analizar si hay diferencias significativas entre los comportamientos de los estudiantes en el curso virtual cuando el curso es totalmente a distancia, se analiza en esta sección el curso de "Herramientas para el Trabajo Colaborativo". Este curso está también estructurado en 4 bloques temáticos, cada uno con su foro de debate y su test on-line y un espacio para entrega de tareas. En este caso no hay indicadores de actividad en sesiones presenciales.

Los algoritmos que se utilizarán para realizar agrupamientos de estudiantes son los algoritmos EM y SimpleK-Means.

A.- Análisis de las actividades virtuales con el algoritmo EM

Utilizando el algoritmo EM fueron identificados 5 grupos de estudiantes, al igual que en el análisis del curso impartido en modalidad B-learning.

En la Figura 6.1 aparecen los 5 grupos creados con los valores medios de los atributos de actividades. El parámetro *log-likelihood* del agrupamiento

es de 36.5217. Como puede verse, el primer clúster agrupa 4 instancias, el segundo 16 instancias, el tercero 6 instancias, el cuarto 6 instancias y el último clúster 25 estudiantes.

Cluster	0	1	2	3	4
N° de Estudiantes	4 (7%)	16 (28%)	6 (11%)	6 (11%)	25 (44%)
T1	0.2761	0 0.2608	0.4214	0	0.2608
T2	0.1515	0	0.3231	0	0.1378
T3	0.3149	0.014	0.4097	0.0476	0.1982
T4	0.4532	0.0588	0.5961	0.3333	0.3318
TR	0.4646	0.0449	0.4696	0.2525	0.2657
FNL	0.5239	0.0713	0.4743	0.4375	0.3351
FGR	1	0	1	0.8	1
FGN	0	0	0.1545	0.2	0
FGL	0.5739	0.0951	0.5925	0.3447	0.5407
F1L	0.5266	0.0353	0.5468	0.22	0.3962
F2L	0	0	0.1545	0	0.167
F3L	0	0	0.1545	0	0
F4L	0.4754	0.0247	0.5142	0.2115	0.2494
F1R	0.4	0.0706	0.4308	0.28	0.35
F2R	0	0	0	0	0
F3R	0.0906	0	0.05	0.0419	0.0419
F4R	0.8786	0	0.8476	0	0.9997
F1N	0.1826	0.0336	0.4709	0.2	0.2208
F2N	0	0	0.2781	0	0.0668
F3N	0.3162	0.0341	0.4719	0.152	0.1767
F4N	0.9385	0	0.8483	0	1

Fig.6.1: Actividades con el algoritmo EM en el sistema E-learning

A la vista de los valores medios de los diferentes atributos en cada grupo, se pueden intuir una serie de características destacables de cada grupo. Además, tomando como referencia esos valores medios y la vinculación con los criterios del modelo AHP, se podría obtener un indicador único de nivel de actividad para cada grupo a partir de la siguiente fórmula:

NA = 0.034*Promedio(FGL,F1L,F2L,F3L,F4L) +
0.132*Promedio(FGR,F1R,F2R,F3R,F4R) +
0.307*Promedio(FGN,F1N,F2N,F3N,F4N) +
0.038*FNL + 0.056*Promedio(T1,T2,T3,T4) + 0.056*TR

Haciendo uso de esta fórmula sobre los centroides de cada grupo, la actividad virtual global de cada uno de ellos es la que se muestra a continuación:

Nivel de actividad virtual del Cluster 0	0,224 Socializadores
Nivel de actividad virtual del Cluster 1	0,017 Desmotivados
Nivel de actividad virtual del Cluster 2	0,2802 Lideres
Nivel de actividad virtual del Cluster 3	0,1049 Medios o gris
Nivel de actividad virtual del Cluster 4	0,2029 Socializadores

Como puede observarse, hay dos grupos que destacan por tener un nivel de actividad virtual más alta que el resto: los clústeres 0 y 2. Si se comparan los valores de los indicadores de actividades sociales en ambos grupos, se observa cómo los estudiantes del cluster 2 tienen valores más altos en los resultados en la participación en los foros (FGN, F1N, F2N, F3N, F4N) y en las pruebas de evaluación virtual de cada bloque temático (T1, T2, T3, T4).

El grupo más numeroso, el clúster 4, corresponde también con el que tiene un nivel de actividad virtual adecuada que en el resto.

A la vista de todos los valores obtenidos se podría decir que las características de los diferentes grupos obtenidos serían las siguientes:

Grupos 0 y 4: Estos dos grupos responden a un perfil muy similar. En el primero se encuentran un 7% de usuarios, mientras que en el segundo está constituido por un 44% de usuarios. Ambos grupos se caracterizan por un

buen nivel de participación en las actividades virtuales, especialmente en los foros. Es por ello que podríamos considerar a estos estudiantes como "**Socializadores"**. Tal vez los de grupo 0 obtienen unos resultados ligeramente mejores que los del grupo 4, pero en cambio, los del grupo 4 parecen estar más pendientes de los mensajes que se publican en los foros, con unos niveles de lectura ligeramente superiores. Estas ligeras diferencias pueden justificar que el algoritmo los haya separado en dos grupos diferentes, pero realmente el perfil de todos ellos es muy similar.

Grupo 1: En este Cluster se encuentra formado por un 28% de usuarios parece que acceden a la plataforma únicamente en momentos puntuales, por ejemplo para realizar tests o para entregar tareas. Prácticamente no participan en los foros temáticos, ni leyendo los mensajes ni haciendo nuevas aportaciones. En definitiva, se podría considerar a este grupo como estudiantes "**Desmotivados**", que no encuentran un suficiente incentivo para participar regularmente en el curso virtual o que no están lo suficientemente implicados en la asignatura.

Grupo 2: Este grupo está formado por un 11% de los estudiantes. Son estudiantes con alta participación en las actividades propuestas a través del curso virtual. En especial son los que más participan en los foros respondiendo a consultas de los compañeros, pero también iniciando nuevos debates. Podríamos decir, por tanto, que responden a un perfil de estudiantes "**Lideres**".

Grupo 3: En este grupo se encuentra igual porcentaje que del grupo 2 con 11% de los estudiantes con actividad creativa en los foros, ya que apenas responden a mensajes o crean hilos de debate nuevos. Su rendimiento académico no se puede considerar malo pero su actividad virtual no es la óptima. Podríamos caracterizar a estos estudiantes como "**Medios o Grises**". La siguiente figura 6.2 muestra el reparto de los estudiantes en los 5 grupos generados por el algoritmo EM y los 4 perfiles de estudiantes que se podrían identificar a partir de las características de cada clúster.

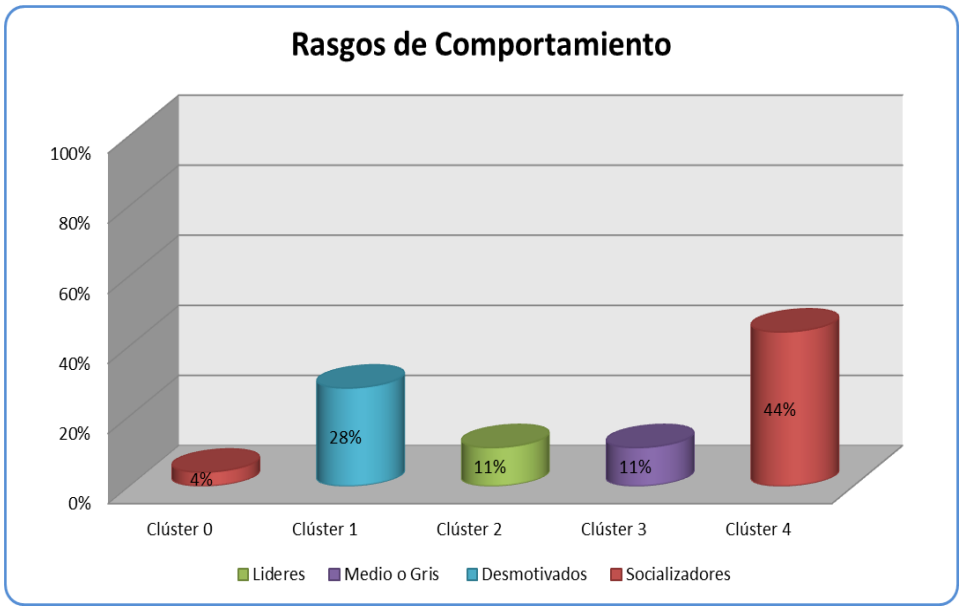


Fig.6.2: Rasgos de estudiantes según actividades con el algoritmo EM en el sistema E-learning

B.- Foros con el algoritmo SimpleK-Means en el sistema E-learning

Aplicando lo mismos análisis en este modelo o sistema e-elarning se plantea la identificación de patrones de comportamiento de los estudiantes en

los foros de debate habilitados en la asignatura, clasificandose en cuatro tipos de usuarios:

4. Inactivos
5. Pasivos
6. Activos
7. Colaborativos

Considerando, la clasificación según el nivel de participación, de menor a mayor participación.

Con objeto de poder identificar esos perfiles entre los estudiantes de la asignatura, se optó por aplicar técnicas de clustering pero utilizando únicamente las variables de participación en los espacios de interacción social (foros) y no aquellas relacionadas con actividades planificadas y programadas por el profesor como son la entrega de tareas, actividades en laboratorios, realización de tests de evaluación. En concreto, los 16 atributos relacionados con la actividad en los foros son:

- Atributos de lectura de mensajes: FNL, FGL, F1L, F2L, F3L, F4L
- Atributos de respuestas a mensajes: FGR, F1R, F2R, F3R, F4R
- Atributos de generación de nuevos hilos: FGN, F1N, F2N, F3N, F4N.

Teniendo en cuenta que en este caso de antemano queremos identificar cuatro grupos de estudiantes, se recurrirá a la utilización de un algoritmo de

clustering sencillo y rápido de ejecución que necesita como parámetro ese número de grupos: el algoritmo Simple k-Means.

Por otro lado, para ordenar los grupos de acuerdo a un nivel global de interactividad social, se utilizarán los pesos obtenidos con el modelo AHP para valorar los diferentes criterios. En concreto, los pesos son:

- Peso de la lectura de mensajes: 0.072
- Peso de las respuestas a otros mensajes: 0,279
- Pesos de la apertura de nuevos debates: 0,649

Con estos pesos, el indicador global de actividad social de cada estudiante o grupo se calcularía con la siguiente fórmula:

AS = 0.072*promedio(FNL,FGL,F1L,F2L,F3L,F4L) +
0.279*promedio(FGR,F1R,F2R,F3R,F4R) +
0.6449*promedio(FGN,F1N,F2N,F3N,F4N)

En la figura 6.3. se muestra el resultado usando el algoritmo SimpleK-Means divide los estudiantes en 4 grupos. En cuanto al reparto de los estudiantes en los grupos, el primero de ellos está formado por 42 instancias (estudiantes), el segundo por 4 instancias, el tercero por 8 instancias, y el último, que además es el menor número con 3 instancias (estudiantes). Tal como se muestra a continuación:

Attribute	Full Data (57)	0 74% (42)	1 7% (4)	2 14% (8)	3 5% (3)
FNL	0.7901	0.7697	0.8389	0.8089	0.961
FGL	0.0769	0.0704	0.121	0.0862	0.0855
FGR	0.0027	0.0024	0.0094	0.0011	0.003
FGN	0.0002	0.0002	0	0	0
F1L	0.0137	0.0129	0.039	0.0056	0.0138
F1R	0.0054	0.004	0.007	0.0096	0.0124
F1N	0.0047	0.0044	0.0062	0.0046	0.0076
F2L	0.0003	0	0	0	0.0059
F2R	0	0	0	0	0
F2N	0.0006	0.0003	0	0	0.0065
F3L	0.0001	0.0001	0	0	0
F3R	0.001	0.0005	0.0052	0.0003	0.0046
F3N	0.1458	0.1449	0.1817	0.1229	0.1718
F4L	0.1494	0.1574	0.1735	0.0933	0.1546
F4R	0.0023	0.002	0.0094	0	0.003
F4N	0.0023	0.002	0.0094	0	0.003

Fig. 6.3: Foros con el algoritmo SimpleK-Means en sistema E-learning

Haciendo uso de esta fórmula sobre los centroides de cada grupo, en foros global de cada uno de ellos es la que se muestra a continuación:

Nivel de actividad virtual del Clúster 0	0,0322 Pasivos
Nivel de actividad virtual del Clúster 1	0,0412 Colaborativos
Nivel de actividad virtual del Clúster 2	0,0290 Inactivos
Nivel de actividad virtual del Clúster 3	0,0403 Activos

Como puede observarse, el grupo que con el nivel de actividad social que destacan es el clúster 1 más alta que el resto: se observa cómo los estudiantes en este grupo tienen valores más altos en los mensajes leídos (FNL,FGL,F1L,F2L,F3L,F4L) y en las respuestas a otros mensajes (FGR,F1R,F2R,F3R,F4R) .

El segundo grupo con actividad social más alta es el correspondiente al Cluster 2, pero con buenos niveles de lectura de mensajes y de colaboración

a través de respuestas a compañeros; sin embargo con poca participación en debate nuevo.

El grupo más numeroso, el Cluster 1, corresponde también con el que tiene un nivel de actividad social más bajo. Los resultados de la participación en los foros en este grupo son también más bajos que en el resto.

A la vista de todos los valores obtenidos se podría decir que las características de los diferentes grupos obtenidos serían las siguientes:

Grupo 0: Este grupo está formado por un 74% de los estudiantes. Son estudiantes con poca participación en los foros respondiendo a consultas de los compañeros, pero nula en nuevos debates. Podríamos decir, por tanto, que responden a un perfil de estudiantes "**Pasivos**".

Grupo 1: En este grupo se encuentra un pequeño porcentaje de los estudiantes (un 7%) con el nivel de actividad social más altos, con buenos niveles de lectura de los mensajes de los foros y de respuesta a mensajes de compañeros, pero sobre todo con poco nivel en el planteamiento de nuevos hilos de debate. Este grupo corresponde, por tanto al perfil de estudiantes "**Colaborativos"**.

Grupo 2: En este Cluster se encuentra el grupo más numeroso de estudiantes, con un 14% de los inscritos en la asignatura, participación medio en las actividades sociales, por la poca actividad creativa en los foros, ya que

apenas responden a mensajes o crean hilos de debate nuevos. Podríamos caracterizar a estos estudiantes como "**Inactivos**".

Grupo 3: En este grupo constituye un 5% estudiante con actividad social alta, Es por ello que podríamos considerar a este pequeño grupo de estudiantes como "**Activos**" pero con buenos niveles de lectura de mensajes y de colaboración a través de respuestas a compañeros; sin embargo no han planteado ningún debate nuevo.

La siguiente figura 6.4 muestra el reparto de los estudiantes en los 4 grupos generados por el algoritmo Simplek-Means y los 4 perfiles de estudiantes que se podrían identificar a partir de las características de cada cluster.

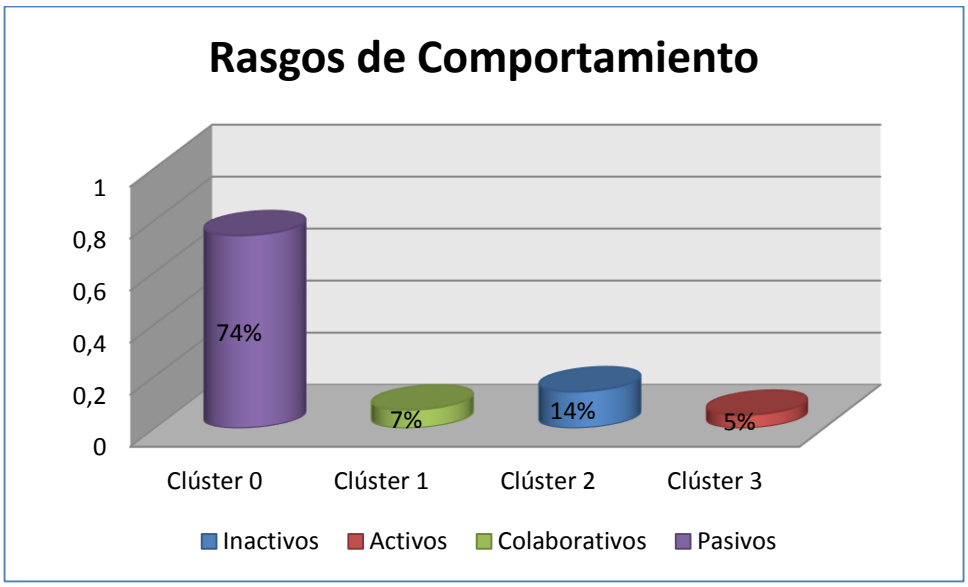


Fig.6.4: Rasgos de comportamiento de estudiantes según su nivel de participación en los foros.

C.- Foros quincenales con el algoritmo SimpleK-Means en el sistema E-learning

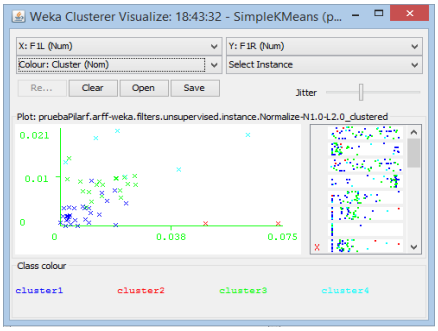
1.- Primera Quincena

```
                         Cluster#
Attribute    Full Data          0          1          2          3
                  (57)       (29)       (2)       (20)        (6)
=============================================================================
F1L             0.0137     0.0068     0.0624     0.0154     0.0257
F1R             0.0054     0.0022          0     0.0076     0.0161
F1N             0.0047     0.0012     0.0125     0.0091      0.005


Clustered Instances

0       29 ( 51%)
1        2 (  4%)
2       20 ( 35%)
3        6 ( 11%)
```

Foros en la 1era quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL1	0,0068	0,0624	0,0154	0,0257
FR1	0,0022	0	0,0076	0,0161
FN1	0,0012	0,0125	0,0091	0,005
AS(1)	0,0019	0,0126	0,0091	0,0096
Perfil	Inactivos	Colaborativos	Pasivos	Activos

Actividad social y perfil en la 1era quincena

En esta 1era quincena en modelo totalmente virtual, el grupo de más alto nivel de actividad social son estudiantes **Colaborativos** es el correspondiente al Cluster 1, que agrupa a 2 estudiantes (4%), con

participaron en los mensajes leídos y nuevos debates, pero nulos en los mensajes respondidos. En un segundo nivel de actividad estudiantes **Activos** se encuentra el Cluster 3, con 6 estudiantes (11%) con buenos niveles de mensajes leídos, respuestas a mensajes de compañeros, y nuevos mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivos**. En este caso se trata del Cluster 2, con 20 estudiantes (un 35% del total) con cierta actividad de lectura de mensajes, respuestas a mensajes de compañeros y actividad de creación de contenido en forma de nuevos mensajes. Igualmente, el grupo con menor índice de actividad global es el Clúster 0 con 29 estudiantes (51 %) considerándose estudiantes **Inactivos** durante esta quincena son estudiantes con un nivel de acceso a los foros leídos muy pequeño, respondidos y en los mensajes nuevos.

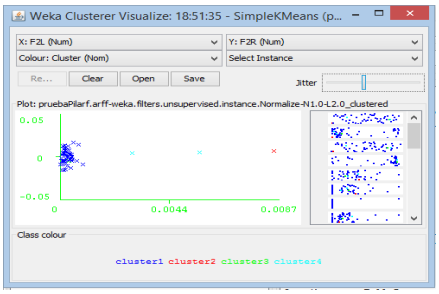
2.- Segunda Quincena

		Cluster#			
Attribute	Full Data	0	1	2	3
	(57)	(53)	(1)	(1)	(2)
F2L	0.0003	0	0.0087	0	0.0045
F2R	0	0	0	0	0
F2N	0.0006	0	0	0.0125	0.0098

Clustered Instances

0	53	(93%)
1	1	(2%)
2	1	(2%)
3	2	(4%)

Foros en la 2da quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL2	0	0,0087	0	0,0045
FR2	0	0	0	0
FN2	0	0	0,0125	0,0098
AS(2)	0,0000	0,0006	0,0081	0,0066
Perfil	Inactivos	Pasivos	Colaborativos	Activos

Actividad social y perfil en la 2da quincena

En esta 2da quincena el grupo con el nivel de actividad social más alto es el Clúster 2, se trata de un grupo pequeño, por un solo estudiantes (un 2% del total) con buenos niveles de planteamiento de nuevos hilos de debate, pero prácticamente nula en la lectura de los mensajes de los foros y de respuesta a mensajes de compañeros. Este corresponde, por tanto al perfil de estudiantes **Colaborativos**. El segundo grupo con actividad social más alta es el correspondiente al Clúster 3, este es por tanto el grupo de estudiantes que responden a un perfil **Activo,** con solo 2 estudiantes (un 4% del total) pero con buenos niveles de lectura de mensajes y debate nuevo; sin embargo no han planteado ninguna colaboración a través de respuestas a compañeros. Siguiendo con el orden de mayor a menor nivel de actividad social, el siguiente grupo correspondería al perfil de estudiante **Pasivo**, en este caso se trata del Clúster 1, con 1estudiantes (un 2% del total) con cierta actividad de lectura de mensajes pero prácticamente nula actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros. Finalmente, como ya se comentó, el grupo con nivel de actividad más bajo es el Cluster 0, con 53 estudiantes (93% del total) que prácticamente no accedieron a los foros. Estos estudiantes son los que responden a un perfil **Inactivo**.

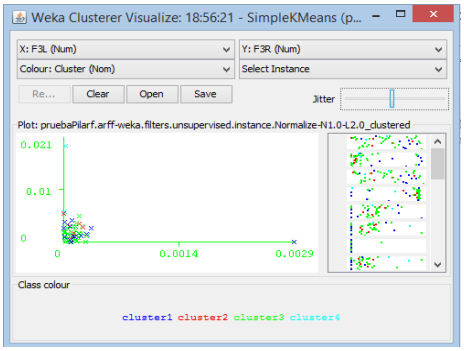
3.- Tercera Quincena

Attribute	Full Data (57)	0 (13)	1 (6)	2 (36)	3 (2)
F3L	0.0001	0.0002	0	0	0
F3R	0.001	0	0.0038	0.0002	0.0144
F3N	0.1458	0.0314	0.2538	0.144	0.5982

Clustered Instances

0	13	(23%)
1	6	(11%)
2	36	(63%)
3	2	(4%)

Foros en la 3era quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL3	0,0002	0	0	0
FR3	0	0,0038	0,0002	0,0144
FN3	0,0314	0,2538	0,144	0,5982
AS(3)	0,0203	0,1647	0,0929	0,3898
Perfil	Inactivos	Activos	Pasivos	Colaborativos

Actividad social y perfil en la 3era quincena

En esta 3era quincena el grupo con el nivel de actividad social más alto es el Clúster 3, con solo 2 estudiantes (un 4% del total) con buenos niveles en el planteamiento de nuevos hilos de debate y de respuesta a mensajes de compañeros., pero prácticamente nula en la lectura de los mensajes de los foros, este corresponde, por tanto al perfil de estudiantes **Colaborativos**. El

segundo grupo con actividad social más alta es el correspondiente al Clúster 1, este es por tanto el grupo de estudiantes que responden a un perfil **Activo,** con solo 6 estudiantes (un 11% del total) pero con buenos niveles de colaboración a través de respuestas a compañeros y debate nuevo; sin embargo no han planteado ninguna lectura de mensajes. El siguiente grupo correspondería al perfil de estudiante **Pasivo**, en este caso se trata del Clúster 2, con 36 estudiantes (un 63% del total) con cierta actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros , pero prácticamente nula actividad de lectura de mensajes. Finalmente, como ya se comentó, el grupo con nivel de actividad más bajo es el Cluster 0, con 13 estudiantes (93% del total) que prácticamente no accedieron a los foros. Estos estudiantes son los que responden a un perfil **Inactivo**.

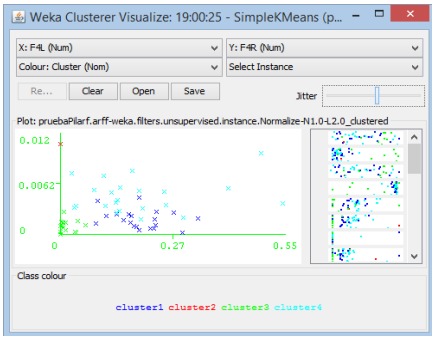
4.- Cuarta Quincena

		Cluster#			
Attribute	Full Data	0	1	2	3
	(57)	(20)	(1)	(14)	(22)
F4L	0.1494	0.1953	0	0.0104	0.2028
F4R	0.0023	0.0008	0.0124	0.0002	0.0045
F4N	0.0023	0.0009	0.0124	0.0002	0.0045

Clustered Instances

0	20	(35%)
1	1	(2%)
2	14	(25%)
3	22	(39%)

Foros en la 4ta quincena



Clústering visual

Atributos	Cluster 0	Cluster 1	Cluster 2	Cluster 3
FL4	0,1953	0	0,0104	0,2028
FR4	0,0008	0,0124	0,0002	0,0045
FN4	0,0009	0,0124	0,0002	0,0045
AS(2)	0,0149	0,0115	0,0009	0,0188
Perfil	Activos	Pasivos	Inactivos	Colaborativos

Actividad social y perfil en la 4ta quincena

En esta 4ta quincena el grupo con el nivel de actividad social más alto es el Cluster 3, con solo 22 estudiantes (un 39% del total) con buenos niveles de la lectura de los mensajes de los foros, en el planteamiento de nuevos hilos de debate y de respuesta a mensajes de compañeros, este corresponde, por tanto al perfil de estudiantes **Colaborativos**. El segundo grupo con actividad social más alta es el correspondiente al Clúster 0, este es por tanto el grupo de estudiantes que responden a un perfil **Activo,** con solo 20 estudiantes (un 35% del total) pero con buenos niveles de lectura de mensajes, colaboración a través de respuestas a compañeros y debate nuevo. El siguiente grupo correspondería al perfil de estudiante **Pasivo**, en este caso se trata del Clúster 1, con 1 estudiantes (un 2% del total) con cierta actividad de creación de contenido en forma de nuevos mensajes o respuestas a mensajes de compañeros , pero prácticamente nula actividad de lectura de mensajes.

Finalmente, como ya se comentó, el grupo con nivel de actividad más bajo es el Cluster 2, con 14 estudiantes (25% del total) que prácticamente poco accedieron a los foros. Estos estudiantes son los que responden a un perfil **Inactivo**.

D.- Resumen de los foros quincenal del algoritmo Simplek-Means en el Sistema E-learning

	Clúster 0	Clúster 1	Clúster 2	Clúster 3
1era quincena	29	20	20	6
	51%	4%	35%	11%
	Inactivos	Colaborativos	Pasivos	Activos
2da quincena	53	1	1	2
	93%	2%	2%	4%
	Inactivos	Pasivos	Colaborativos	Activos
3ra quincena	13	6	36	2
	23%	11%	63%	4%
	Inactivos	Activos	Pasivos	Colaborativos
4ta quincena	20	1	14	22
	35%	2%	25%	39%
	Activos	Pasivos	Inactivos	Colaborativos

Tabla 6.2: Foros en la 4 quincenas con el Algoritmo Simplek-Means en el sistema E-learning

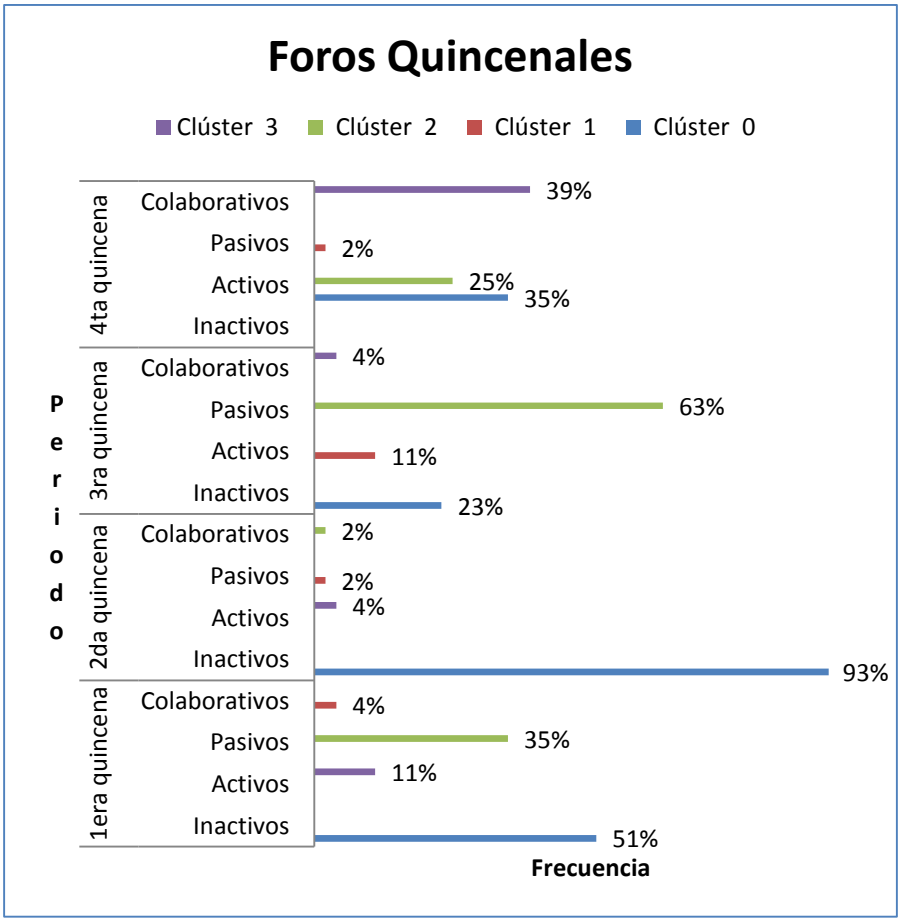


Fig 6.5: Foros en la 4 quincenas con el Algoritmo Simplek-Means

En la figura 6.5 Durante las 4 quincenas, los estudiantes con participación en los foros en el sistema E-learning con el Algoritmo Simplek-Means, se concluye, en la 1era quincena muestra la participación de (51%) estudiantes, con rasgos de comportamiento ***Inactivos,*** pues esto indica que los estudiantes participan con buenos niveles de mensajes leídos se justifica, por el inicio a esta modalidad del curso totalmente virtual, así mismo en esta quincena continuando con el nivel de interactividad social se considera con (35%) estudiantes ***Pasivos***. pues esto indica que los estudiantes pues participan con

bajos niveles de mensajes leídos, mensajes respondidos, pero nulo en los nuevos mensajes.

En la 2da quincena muestra el mayor porcentaje de participación (93% del total) de estudiantes ***Inactivos*** implica que la interactividad social es similarmente que en la 1er quincena. Seguidamente en esta quincena observamos la participación de estudiante (4% del total) estudiantes ***Activos*** pues esto indica que los estudiantes tienen buenos niveles de lectura de mensajes y de colaboración a través de respuestas a compañeros; sin embargo no han planteado ningún debate nuevo.

En la tercera quincena se observa (63% del total) de estudiantes con rasgos de comportamiento ***Pasivos,*** pues esto indica que los estudiantes pues participan con buenos niveles de mensajes leídos, pero nulo a mensajes respondidos, y los nuevos mensajes seguida en esta quincena (23%) de usuarios con rasgos de comportamiento ***Inactivos***. Finalmente en la 4ta quincena, muestra participación de usuarios (39% del total) con rasgos de comportamiento ***Colaborativos***, con el nivel de actividad social más altos, con buenos niveles de lectura de los mensajes de los foros y de respuesta a mensajes de compañeros, pero sobre todo con alto nivel en el planteamiento de nuevos hilos de debate. Seguido del grupo de estudiantes que corresponden a un perfil **Activo.** Se trata de un grupo de (un 35% del total) pero con buenos niveles de lectura de mensajes y de colaboración a través de respuestas a compañeros; sin embargo no han planteado ningún debate nuevo.

4.4. COMPARACIÓN DE LOS MODELOS B-LEARNING Y LA E-LEARNING

En esta fase se comparara tanto en los modelos E-learning, y B-learning, los rasgos de comportamiento evaluados con el proceso de análisis jerárquico y la minería de datos los resultados que se muestra en las tablas 7.1 y 7.2

A.- Comparación de las actividades entre los modelos B-learning y E-learning

Modelo	B-learning				
Clúster	0	1	2	3	4
Instancias	16	24	110	7	154
% de instancias	5%	8%	35%	2%	50%
Perfil	Socializadores	Medio o Gris	Medio o Gris	Lideres	Desmotivados
Modelo	E-learning				
Clúster	0	1	2	3	4
Instancias	7	16	6	6	25
% de instancias	4%	28%	11%	11%	44%
Rasgos de comportamiento	Lideres	Desmotivados	Socializadores	Medios o Gris	Lideres

Tabla 7.1: Comparación de las actividades entre B-learning y E-learning

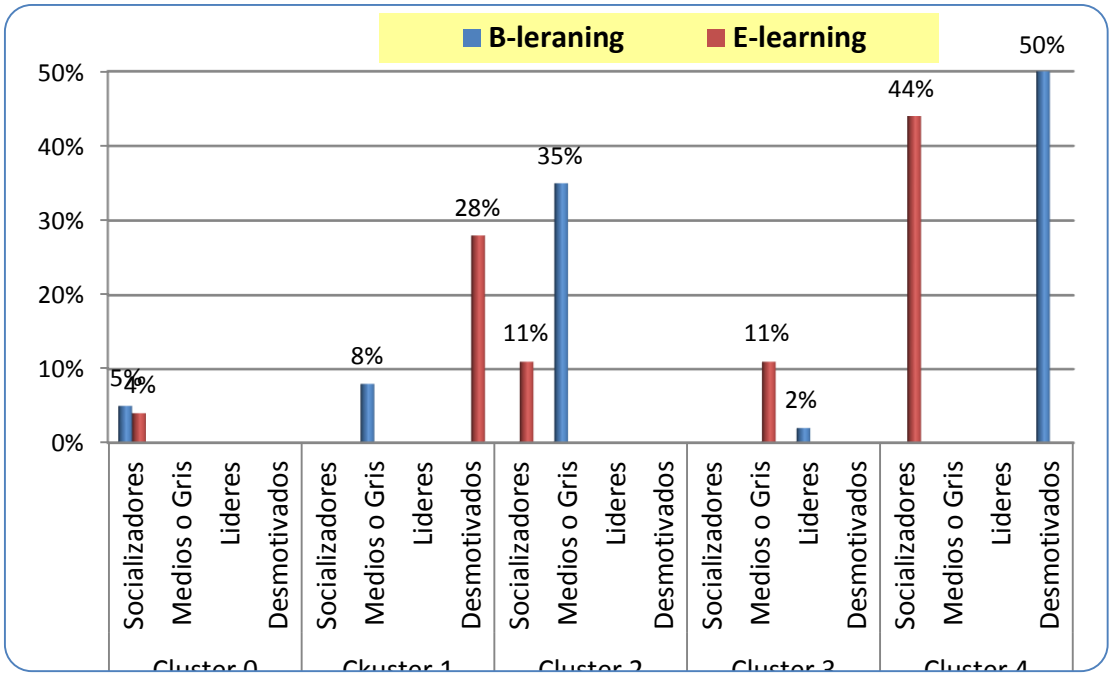


Fig. 7.1: Comparación de las actividades entre los modelos B-learning y E-learning

En la fig. 7.1, observamos la participación de las actividades realizadas en los modelos B-learning y E-learning de acuerdo a los grupos formados de estudiantes es así que en el Cluster 0 el rasgos de comportamiento son similares los estudiantes **Socializadores** en estos modelos, pues en el Cluster 1 tiene mayor valor significativo el modelo E-learning con rasgo de comportamiento desmotivados, mientras que en el modelo B-learning destaca el rasgo de comportamiento **Medio o Gris,** continuando con el siguiente grupo Cluster 2, el valor significativo está representado por el modelo E-learning con el rasgo de comportamiento de estudiantes **Medio o Gris,** seguido del rasgo de comportamiento de los estudiantes **Socializadores** ,pero en el modelo B-learning, siguiendo con los rasgos de comportamiento de los estudiantes, en el Cluster 3 observamos el más destacado es el modelo E-learning con comportamiento **Medio o Gris,** mientras que en modelos B-

learning destacan los estudiantes **Lideres.** Finalmente en el Cluster 4 el más significativo es el modelo E-learning con el rasgo de comportamiento de estudiantes **Socializadorees**, dado que en el modelo B-learning destacaron los estudiantes **Desmotivadores**

b.- Comparación de los foros entre los modelos B-learning y E-learning

Modelos o Sistemas	B-learning			
Clúster	0	1	2	3
Instancias	41	249	48	4
% de instancias	12%	73%	14%	1%
Rasgos de comportamiento	Pasivos	Inactivos	Activos	Colaborativos
Modelos o Sistemas	E-learning			
Clúster	0	1	2	3
Instancias	42	4	8	3
% de instancias	74%	7%	14%	5%
Rasgos de comportamiento	Pasivos	Colaborativos	Inactivos	Activos

Tabla 7.2: Comparación de los foros entre B-learning y E-learning

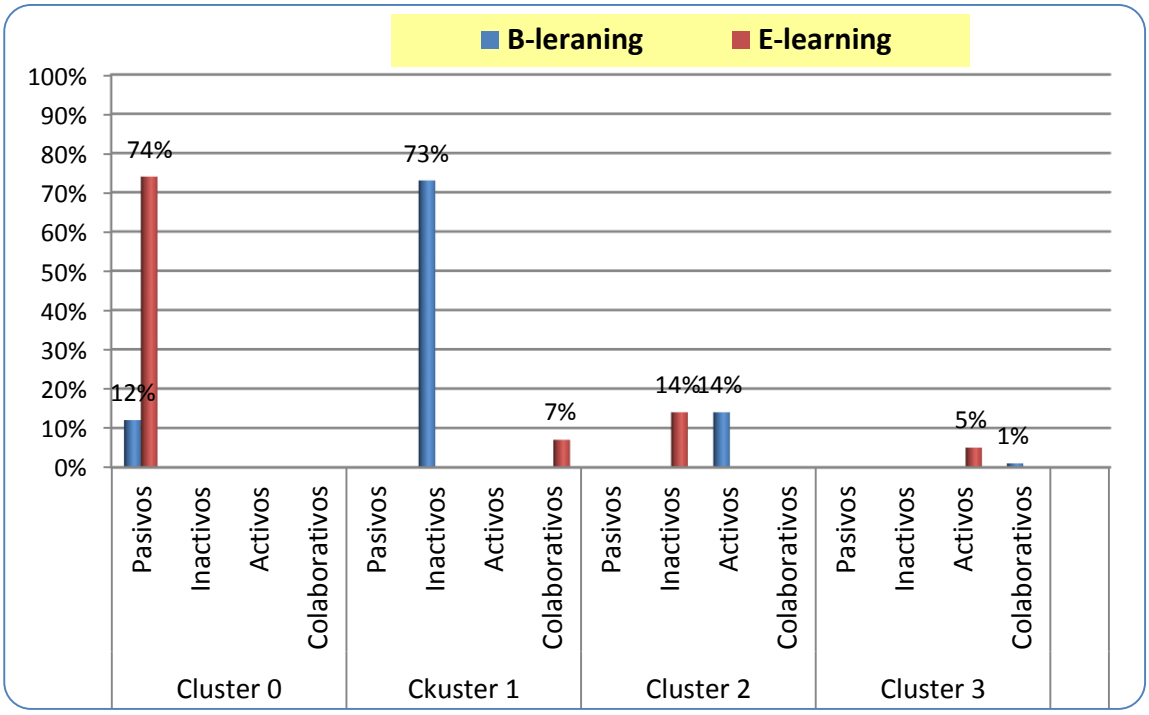


Fig.7.2: Comparación de los foros entre B-learning y E-learning

En la Fig 7.2, observamos la participación de los foros realizadas en los modelos B-learning y E-learning de acuerdo a los grupos formados de estudiantes, en el Cluster 0 ambos modelos son representados por estudiantes **Pasivos,** pero el más significativos es el modelo E-learning con el respecto al modelo B-learning. En el cluster 1 el más significativo es el modelo B-learning, con rasgo de comportamientos de estudiantes **Inactivos,** mientras que en el modelo E-learning destacan los estudiantes **Colaborativos,** siguiendo con la comparación de comportamiento estos modelos, observamos en el Cluster 2 que ambos modelo son iguales en representación pero en rasgos de comportamiento en el modelo E-learning son estudiantes **Inactivos**, mientras que en el modelo B-learning son **Activos.** Finalmente en el Cluster 3 el más significativos es el modelo E-learning con rasgo de comportamiento de

los estudiantes **Activos,** mientras que en el modelo B-learning destacan el rasgo de comportamiento de estudiantes **Colaborativos**

V. DISCUSIÓN

En este trabajo se han utilizado técnicas de minería de datos para identificar grupos de estudiantes de acuerdo a su patrón de comportamiento y su rendimiento así como el acceso a los foros. Se ha observado un predominio de estudiantes que adoptan posiciones más pasivas, pero que sin embargo no necesariamente obtienen bajos rendimientos académicos. Sí que es cierto que aquellos estudiantes que están más predispuestos a la colaboración por lo general obtienen mejores resultados

Lograr la implicación de todo el estudiantado en las iniciativas de los modelos E-learning, B-learning, requiere implementar prácticas pedagógicas que favorezcan no solo entornos de aprendizaje personalizados sino que se centre en las expectativas e intereses de los estudiantes, logrando una mayor interactividad e implicación de los estudiantes Inactivos y Pasivos en este proceso. En este sentido, la generación de conocimiento mediante análisis automático de datos puede servir para el diseño de estrategias de aprendizaje personalizadas y desarrollar sistemas de autorización inteligente

De esta manera, el personal docente tiene un modelo que puede facilitar individualización de la enseñanza a las características particulares presentado por cada uno de los estudiantes en el LMS. (Limayem y Cheung ,2011) demostró la importancia de los profesores para promover el uso sostenible y prolongado de basado en Internet tecnologías en contextos universitarios de aprendizaje. Además, los datos se pueden obtener que mejora la toma de decisiones en los procesos de la educación formativa, destinada a la reorientación de aquellos aspectos que han demostrado ser más problemático o no lo hacen responder a los objetivos definidos a priori.

Este significa que, al igual que en el caso de los procesos de formación fuera de línea, no existe una relación de enseñanza y aprendizaje lineal y no se puede prever con precisión el desarrollo y los resultados de los procesos de B-learning y E-learning programado al mismo tiempo, estas diferencias de rendimiento de los estudiantes revelan que a fin de comprender lo que sucede en las experiencias de E-learning es necesario para ver a los participantes con más interactividad.

Por último, esta investigación nos ha permitido afirmar que los patrones de comportamiento del estudiante en los modelos E-learning y B-learning son muy disparejos. Por lo tanto, los estudiantes pueden mostrar diferentes tipos de comportamiento que van de la participación permanente en el curso virtual. Además, el uso de LMS para la gestión de los procesos de aprendizaje requiere voluntad por parte del estudiante y series de habilidades que no siempre se logran a un grado óptimo. En este sentido, estudios recientes demuestran la necesidad a la investigación de la forma en

que los estudiantes utilizan este tipo de herramientas (Lust et al. 2012). Con el fin de maximizar su potencial, sería necesario diseñar estrategias metodológicas que promover la participación activa de los estudiantes y la colaboración construcción del aprendizaje.

VI. CONCLUSIONES

1.- El Proceso de Análisis Jerárquico y las técnicas de minería de datos ha permitido obtener el rasgo de comportamiento de los estudiantes en las modalidades B-learning y E-learning, utilizando los pesos generados con AHP, permitiendo identificar con mucha facilidad los rasgos característicos de cada grupo.

2.- Dentro del campo de la educación, el área de aplicación de minería de datos en los modelos e-learnig, B-learning resulta muy útil y de interés para los profesores para descubrir patrones de comportamiento de sus alumnos así como para tener información objetiva sobre del uso de sus cursos.

3.- El modelo AHP nos ha permito evaluar el nivel de interactividad de los alumnos, permite a uno identificar un índice cuantitativo de la interactividad del estudiante que puede variar en términos de la

relevancia que cada profesor da a los diferentes indicadores utilizados para su construcción.

4.- La herramienta utilizada (Expert Choice) y Weka para realizar el estudio de comportamiento de los estudiantes en las modalidades E-learning y el b-learnig ha resultado ser eficiente y rápida.

VII. RECOMENDACIONES

1.- Para cuantificar la efectividad del proceso de enseñanza-aprendizaje, usar **Minería de Datos** en los sistemas educativos por ser muy apropiada para descubrir información "escondida" en las bases de datos. Los métodos pueden ser aplicados para explorar, visualizar, y analizar datos con la finalidad de identificar patrones útiles aplicables a la evaluación de la actividad del usuario y descubrir más profundamente como aprenden los estudiantes.

2.- Se recomienda trabajar con los centroides de los atributos obtenidos con SimpleK-Means, que fue el que presentó una mejor clasificación en su aplicación, ya que los Cluster es obtenidos mediante este algoritmo presentaron mayor consistencia y similitud entre sus características. Los valores de estos centroides serán los que se consideren para la definición de los rasgos de comportamiento de los estudiantes.

3- Aplicar nuevas tecnologías en los procesos educativos puesto que se modifican las condiciones de la enseñanza y el aprendizaje. Los roles y las habilidades tradicionales del estudiante y del docente deben cambiar. Los programas de educación basados en Web requieren un mayor esfuerzo, implicación y habilidades para gestionar información.

4- Para que el profesor realice valoraciones comparativas debe usar la Metodología AHP demuestra ser muy apropiada, mediante el escala discreta 1-9 diseñada por Saaty

VIII. REFERENCIAS BIBLIOGRÁFICAS

[1] Anaya, A.R. and Boticario, F.G. 2011. Towards improvements on domain independent measurements for collaborative assessment Proceedings. 4th International Conference on Educational Data Mining EDM 2011. Eindhoven 317–318.

.

[2] Baker, R. S., 2011. La minería de datos para la educación. En: B. McGaw, E. Baker, y P. Peterson, eds. Enciclopedia Internacional de educación. 3rd ed. Vol. 7. Oxford: Elsevier, 112–118.

[3] Bodea, CN, Dascalu, MI, y Lytras, MD, 2012. Un recomendador motor para la retroalimentación personalizada avanzada en E-learning entornos. Revista Internacional de Ingeniería Educación. 28 (6), 1326–1333

[4] Brusilovsky, P. 2003 “Adaptive and Intelligent Web-based Educational Systems”, International Journal of Artificial Intelligence in Education, , IO Press.

[5] Brusilovsky, P., Schwarz, E., & Weber, G., “ELM-ART: An intelligent tutoring system on World Wide Web”, Third International Conference on Intelligent Tutoring Systems, ITS-96 Berlín: Springer Verlag.

[6] Cobo, A; Rocha, R; Álvarez, 2012. Análisis y Evaluación del Rendimiento Académico y Patrones de Comportamiento de los estudiantes en un entorno Interactivos B-learning Utilizando Técnicas De Minería De Datos. TBIT853838

[7] De Wever, B., et al., 2010. Roles as a structuring tool in online discussion groups: the differential impact of different roles on social knowledge construction. Computers in Human Behavior, 26 (4), 516–523.

[8] Dringus, L.P. y Ellis, T., 2005. Usando minería de datos como una estrategia para la evaluación de los foros de discusión asincrónica. Informática e Educación.

[9] Garcia, A. and Suárez. C., 2011. Interacción virtual y aprendizaje cooperativo. Un estudio cualitativo. Revista de Educación, 354, 473–498.

[10] García, M., & Quintales, L. M. 2002. Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. Salamanca.

[11] Garre, M., Cuadrado, J. C., & Sicilia, M. (2005). Comparación de diferentes algoritmos de Clustering en la estimación de coste en el desarrollo de software. Universidad de Alcalá, Departamento de Ciencias de la Computación, Alcalá de Henares, Madrid.

[12] Han, J. y Kamber, M., 2006. La minería de datos. Conceptos y técnicas. Morgan Kaufmann Publishers.

[13] Ishizaka, A. y Lusti, M., 2006.Como derivan las prioridades en AHP: un estudio comparativo. Diario Central Europeo de Operaciones Investigación, 14 (4), 387-400.

[14] Itmazi, J.A.S., "Sistema Flexible de gestión del E-learning para soportar el aprendizaje en las universidades tradicionales y abiertas". PhD Thesis. University of Granada, Spain.

[15] Johnson, M., et al. , 2011. Herramienta EDM. Actas de la cuarta conferencia internacional sobre minería de datos educativos, 349-350.

[16] Liu, P., et al. , 2009. E -learning Evaluación plataforma usando fuzzy AHP. Conferencia internacional sobre inteligencia computacional y la ingeniería de software.

[17] Paquete de minería de datos Weka. Disponible en Abril de 2008 en http://www.cs.waikato.ac.nz/ml/weka/.

[18] Romero, C. y Ventura, S., 2010. La minería de datos para la Educación: una revisión del estado de la técnica. IEEE Transactions on Systems, Hombre y Cibernética, Parte C.

[19] Romero, C., Ventura, S., y García, E., 2007. La minería de datos en sistemas de gestión de cursos Moodle: estudio de casos y de tutoría Informática y Educación. 51 (1), 368-384

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

[21] Srivastava, J.; Mobasher, B.; Cooley, R., "Automatic Personalization Based on Web Usage Mining", Communications of the Association of Computing Machinery, 2000.

[22] Saaty, T. L., 1980. El proceso de jerarquía analítica: la planificación, establecimiento de prioridades, la asignación de recursos. Nueva York: McGraw-Hill.

[23] Siemens, G., Baker RSjd (2012). "Analíticas de aprendizaje y enseñanza de minería de datos: hacia la comunicación y la colaboración". Actas de la 2ª Conferencia Internacional sobre Aprendizaje y Conocimiento Analíticas: 252-254.

[24] Talavera, L. y Gaudioso, E., 2004. Minería Datos caracterizar los Clúster con comportamientos similares en colaboración no estructurada espacios. Taller sobre inteligencia artificial en CSCL, Conferencia Europea 16a en la inteligencia artificial.

[25] Tzeng, G., Chiang, C., and Li, C., 2007. Evaluating intertwined effects in E-learning programs: a novel hybrid MCDM model based on factor analysis and DEMATEL. Expert Systems with Applications, 32 (4), 1028–1044.