MINERÍA DE DATOS EDUCATIVOS EN PLATAFORMAS DE TELEFORMACIÓN

Espigares, Manuel Jesús; García, Rafael y Quiñones, Carlos J.
Universidad de Sevilla

1. INTRODUCCIÓN

Esta comunicación aborda una nueva metodología de evaluación aplicable en los procesos educativos más característicos de la Sociedad del Conocimiento. El aprendizaje regulado mediante plataformas de teleformación (Educational Management System -EMS-) se sitúa en esta sociedad en un primer plano. Nos referimos a las técnicas estadísticas de la minería de datos, que aplicadas en este campo se denominan Minería de Datos Educativos (Educational Data Mining –EDM-). Éstas, implican la generación de modelos estadísticos de carácter correlacional y predictivo (exploratorias y confirmatorias) con los datos almacenados (registros de las acciones) en dichas plataformas. En esta línea, nuestro trabajo plantea una manera organizada de evaluar distintos objetos de aprendizaje online y dinámicas del proceso de aprendizaje registrado en las EMS. Como ejemplo, la aplicamos al estudio de un modelo educativo-musical que emplea una plataforma Moodle (Espigares y García, 2006) (ver. 6.0) en centros TIC de secundaria (www.grupodime.es / plataforma educativa del IES La Campana). Los resultados de este estudio con EDM y su utilidad para la toma de decisiones educativas constituyen el debate principal cara a la valoración metodológica de estas técnicas. Como resumen básico podemos decir que modelos estadísticos generados con la aplicación de la minería de datos educativos, reflejan la necesidad de rediseñar la estructura de las plataformas, de forma que ahorremos tiempo y esfuerzo en la exploración y el aprendizaje de la navegación, tratando de hacer el diseño de las plataformas “más transparente” y menos tortuoso para el alumnado; todo ello, con el objetivo de optimizar el aprendizaje y facilitar un mayor aprovechamiento y mejorar la eficacia educativa.

2. MATERIAL Y MÉTODO

2.1. Muestra de datos

En nuestra investigación se trabaja con una muestra de alumnado de 250 sujetos-usuarios/as, oficialmente matriculados en el centro de la aplicación y asistentes a clases. La distribución por sexos varía según la herramienta de recogida de información elegida. Nos encontramos ante una población con diversidad de niveles socioeconómicos de procedencia, que cursa la Educación Secundaria Obligatoria y con edades comprendidas entre 12 -16 años.

La plataforma educativa con la que trabaja el alumnado en nuestro Modelo educativo almacena casi 50.000 registros y alrededor de 100 variables de estudio diferentes a las que sometemos en primer lugar a una limpieza, selección de la información que nos interesa someter a análisis y la aplicación de descriptivos básicos de frecuencias y modelos estadísticos del campo de la minería de datos educativos. Ante una cantidad de información tan grande el software estadístico tradicional se ve limitado y no posee la potencia necesaria para analizar de forma pormenorizada y exhaustiva la información. Dicha circunstancia nos lleva a plantear la necesidad del uso de un tipo de software específico para este tipo de tareas que nos permita trabajar con comodidad y precisión, de forma fiable y válida con los datos educativos con los que contamos.

2.2. Metodología

Se utiliza una metodología de investigación híbrida o mixta, haciendo uso de herramientas de recogida de información cualitativa y cuantitativa. Este estudio emplea diferentes herramientas de recogida de información como escalas Likert sobre trabajo en la plataforma, escalas de actitud hacia la plataforma y escalas de aprendizaje con la plataforma así como modelos estadísticos a partir de los datos registrados en la plataforma educativa. Además, utilizamos sistemas de categorías “abductivos” para analizar los datos de un grupo de discusión formado por alumnado de distintos cursos y grados de aprendizaje así como de motivación de cara al empleo del modelo pedagógico Bordón, todos ellos registrados en video.

2.3. Software empleado en el proceso de investigación

En nuestro trabajo con minería de datos educativos empleamos de forma básica tres herramientas informáticas que son SPSS Clementine (versión 11.1), Weka (versión 1.4) y Statistica (versión 8.0). Los dos primeros programas los empleamos para la modelización estadística y el último posee posibilidades de configuración para gráficos tridimensionales que favorecen la visualización de los datos con los que trabajamos.

2.4. El rol de la minería de datos educativos en el proceso de evaluación del aprendizaje online

En la actualidad en la minería de datos educativos hay dos líneas de trabajo, la primera de ellas es la reingeniería informática de algoritmos para la minería de datos (Romero et alter, 2008) y la segunda, en la que ubicamos nuestro trabajo, plantea la implementación de técnicas de la minería de datos al campo de la educación y la evaluación de aprendizajes online.

Nuestro estudio plantea la evaluación del diseño, el proceso y el impacto de un modelo educativo-musical (que denominamos *Modelo Bordón*) basado en las TIC. Para ello,

empleamos distintas herramientas tanto de corte cualitativo como cuantitativo a través de una metodología integrada. Dentro de la evaluación del proceso de nuestro modelo planteamos de forma novedosa en el campo de los métodos de investigación educativa el uso de técnicas estadísticas basadas en la minería de datos. La minería de datos procede del campo del marketing, la banca y los negocios y es empleada para analizar grandes volúmenes de información registrada en tiempo real en bases de datos. La aplicación de estas técnicas supone la generación de distintos modelos estadísticos descriptivos y predictivos a partir de los datos que quedan registrados y almacenados en esas plataformas en tablas con registros y variables. La tarea que realizamos consiste en limpiar y seleccionar las variables significativas para el estudio y confeccionar modelos estadísticos con los datos con los que contamos. A través de esta forma de trabajar con los datos educativos conseguimos darle la importancia y el significado que esa ingente cantidad de información registrada tiene y nos permite extraer conocimiento útil sobre el proceso de uso de una herramienta telemática para la formación musical en red.

3. ANÁLISIS

En este apartado hacemos una evaluación cuantitativa de la actividad generada en la plataforma educativa musical. Esta tarea de evaluación sigue un método que denominamos: *Minerización de datos educativos musicales generados en plataformas online.* Para ello, contamos con herramientas de la estadística clásica y la minería de datos educativos. En cuanto a la estadística clásica hacemos uso de análisis descriptivos de frecuencias. La minerización educativa o minería educativa de datos, consiste en la aplicación de modelos estadísticos descriptivos y predictivos orientados a la extracción de patrones de comportamiento significativos de las *tablas que registran toda la actividad* en las plataformas educativas[1]. La minerización abarca múltiples técnicas que van desde la limpieza y selección de variables, la reducción de la dimensionalidad, la segmentación, la clasificación, las reglas de asociación y opciones visualización de los datos, para completar este método incluimos procedimientos de minerización complementarios de cualquier tipo. Queremos destacar que el método que seguimos para trabajar con datos educativos online, es transportable a cualquier otra materia distinta de la música, o nivel diferente de la Educación Secundaria Obligatoria, por lo que la utilidad de nuestro método radica en su compatibilidad y portabilidad.

3.1. Análisis estadísticos descriptivos básicos

a) *Actividad en la plataforma del profesor-investigador y el alumnado:* El 20,2 por ciento (9.934 registros) de la actividad de la plataforma es la realizada por el profesor y el 79,8 por ciento (39.161 registros) por el total del alumnado, es decir, una cuarta parte, aproximadamente de la actividad global. A continuación mostramos estos datos en una tabla y un gráfico de barras. El número total de registros de las tablas de la plataforma educativa es de 49.095, contando con la actividad del profesor y del alumnado.

[1] El término KDD significa en inglés Knowledge Databases Discovery (extracción de conocimiento [útil] partir de bases de datos).

		Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje acumulado
Validos	Profesor	9934	20,2	20,2	20,2
	Alumnado	39161	79,8	79,8	100,0
	Total	49095	100,0	100,0	

Tabla 1. Frecuencia de participación en la plataforma del profesor y el alumnado.

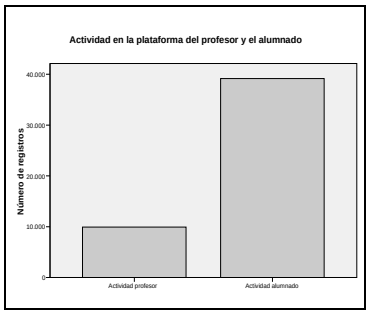


Figura 1. Actividad del profesor y del alumnado.

b) *Actividad en la plataforma por cursos (profesor-investigador y alumnado):* En primero el porcentaje participación del 20,8 por ciento (10.182 registros), en segundo el 27,9 por ciento (13.362), en tercero el 38,6 por ciento (18.914) y en cuarto el 12,8 (6.270 registros). A continuación mostramos estos datos en una tabla y un gráfico de barras.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Validos	1º	10182	20,8	20,8	20,8
	2º	13662	27,9	27,9	48,6
	3º	18914	38,6	38,6	87,2
	4º	6270	12,8	12,8	100,0
	Total	49028	100,0	100,0	

Tabla 2. Actividad de la plataforma educativo-musical distribuida por cursos.

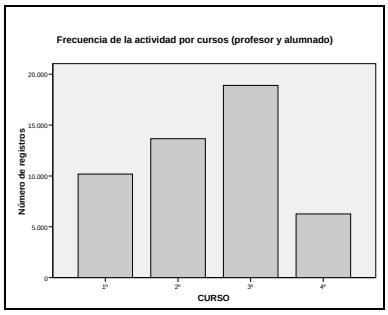


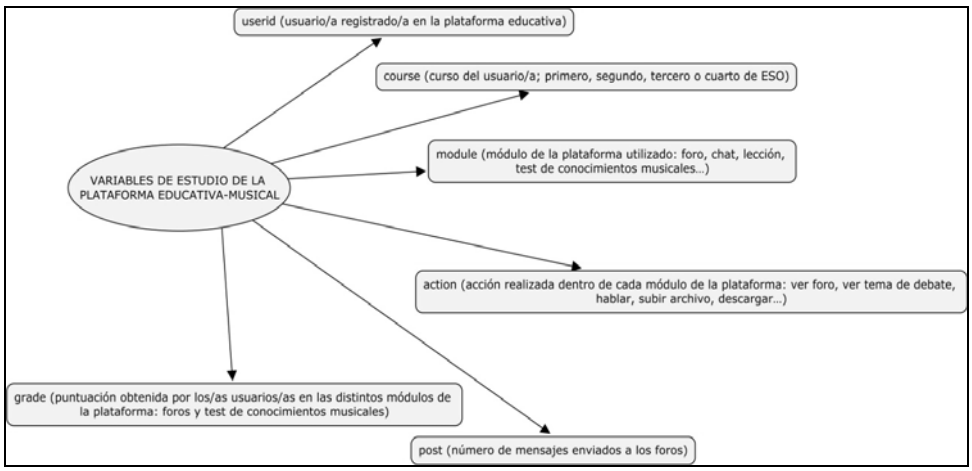
Figura 2. Gráfico de barras con la actividad de la plataforma por cursos.

Tal y como hemos comprobado en los análisis, la cantidad de registros y datos educativos es tan numerosa que optamos por mostrar los resultados de los modelos así como de los procedimientos de minería musical que hemos aplicado, debido a que las tablas con los datos (49.095 registros), excederían los límites razonables de extensión y espacio de este

trabajo en papel. Debido a esto, este trabajo está acompañado del CD con todos los datos recopilados de las tablas de registro de actividad de la plataforma educativa-musical.

3.2. Limpieza de la información: reducción y selección de variables

Tras la tarea de exploración de las tablas de la plataforma educativa, concluimos que aquellas variables que nos arrojan una información más valiosa en nuestra investigación son: *userid* (código numérico que identifica al alumnado participante en la plataforma educativa), *course* (curso del usuario/a, primero, segundo, tercero o cuarto de ESO), *module* (módulo de la plataforma utilizado: foro, chat, lección, cuestionario…), *action* (acción realizada dentro de cada módulo de la plataforma: ver, subir archivo, descargar…), *post* (número de mensajes enviados a los foros), *quiz* (realizar test de conocimientos musicales), *grade* (puntuación obtenida por los usuarios/as en las distintos módulos de la plataforma: foros, lecciones o cuestionarios). A continuación mostramos en un mapa conceptual las variables de estudio de la plataforma educativa-musical:



Mapa conceptual 1. Variables de estudio de la plataforma educativa-musical.

3.3. Generación de modelos estadísticos con la minería de datos educativos

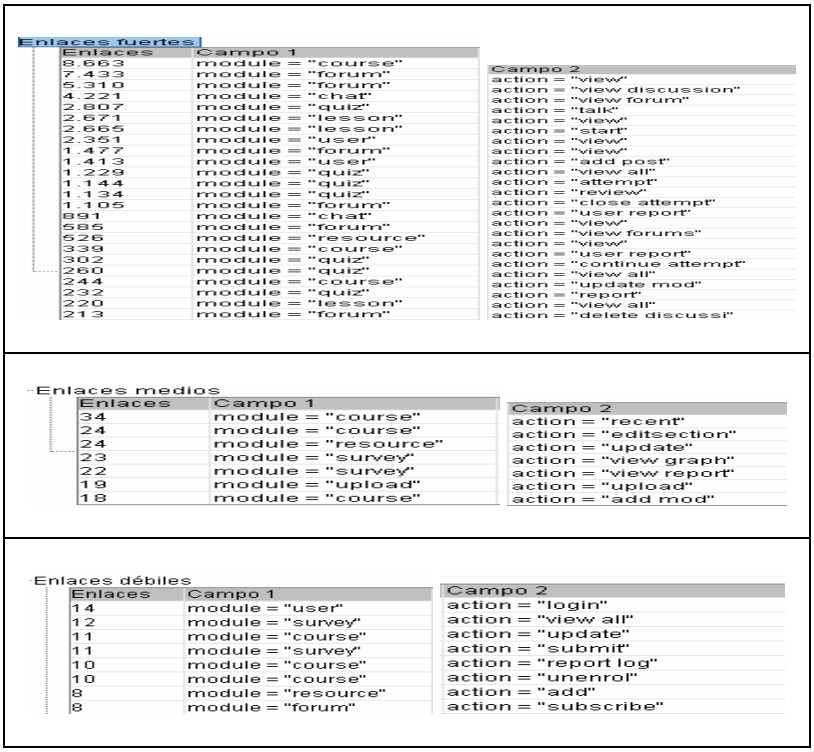
Las tareas de minerización (término tomado de Sierra, 2006) musical que planteamos, consisten en la extracción de patrones de comportamiento de la información registrada en la base de datos de nuestra plataforma educativa. Dichas tareas, se basan en aprehender conocimiento útil y significativo (partiendo de los datos almacenados en las tablas de registro de actividad de nuestro soporte informático), mediante la aplicación de modelos estadísticos variados. En nuestro trabajo, las posibilidades de aplicación del *Data mining* en la evaluación de aprendizajes online son muy variadas ya que abarcan múltiples técnicas exploratorias, gráficas y clasificatorias.

En cuanto al tiempo invertido en la generación de los modelos, en ningún caso supera los 5 segundos, lo cual nos posibilita una gran cantidad de combinaciones y variaciones en el análisis hasta alcanzar un modelo óptimo que ofrezca la máxima validez y fiabilidad.

De todos los modelos estadísticos que confeccionamos en nuestra investigación, por motivos de espacio mostramos sólo el de *malla direccional* para medir el grado de correlación entre variables.

Los *nodos de malla direccional* se utilizan para ilustrar la fuerza de las relaciones existentes entre los valores de dos o más campos simbólicos. Las conexiones se muestran en un gráfico con distintos tipos de líneas para indicar conexiones de creciente fuerza. En nuestro caso medimos el grado de correlación entre dos variables: *módulo* y *acción*. Las líneas más gruesas indican un índice mayor de relación y las más débiles un menor índice de relación. La acción "ver curso" destaca por encima de las demás con 8.663 enlaces, seguida por "ver foro de discusión", con 7.433 enlaces y "hablar en el chat". Las dos primeras acciones denotan acciones de tipo exploratorio y la última muestra una actividad de aprendizaje.

A continuación mostramos las tablas con los índices de enlaces entre los módulos de la plataforma y las acciones realizadas.



Enlaces fuertes

Enlaces	Campo 1	Campo 2
8.663	module = "course"	action = "view"
7.433	module = "forum"	action = "view discussion"
5.310	module = "forum"	action = "view forum"
4.221	module = "chat"	action = "talk"
2.807	module = "quiz"	action = "view"
2.671	module = "lesson"	action = "start"
2.665	module = "lesson"	action = "view"
2.351	module = "user"	action = "view"
1.477	module = "forum"	action = "add post"
1.413	module = "user"	action = "view all"
1.229	module = "quiz"	action = "attempt"
1.144	module = "quiz"	action = "review"
1.134	module = "quiz"	action = "close attempt"
1.105	module = "forum"	action = "user report"
891	module = "chat"	action = "view"
585	module = "forum"	action = "view forums"
526	module = "resource"	action = "view"
339	module = "course"	action = "user report"
302	module = "quiz"	action = "continue attempt"
260	module = "quiz"	action = "view all"
244	module = "course"	action = "update mod"
232	module = "quiz"	action = "report"
220	module = "lesson"	action = "view all"
213	module = "forum"	action = "delete discussi"

Enlaces medios

Enlaces	Campo 1	Campo 2
34	module = "course"	action = "recent"
24	module = "course"	action = "editsection"
24	module = "resource"	action = "update"
23	module = "survey"	action = "view graph"
22	module = "survey"	action = "view report"
19	module = "upload"	action = "upload"
18	module = "course"	action = "add mod"

Enlaces débiles

Enlaces	Campo 1	Campo 2
14	module = "user"	action = "login"
12	module = "survey"	action = "view all"
11	module = "course"	action = "update"
11	module = "survey"	action = "submit"
10	module = "course"	action = "report log"
10	module = "course"	action = "unenrol"
8	module = "resource"	action = "add"
8	module = "forum"	action = "subscribe"

Tabla 3. Enlaces fuertes, medios y débiles de las variables module y action.

A continuación mostramos el gráfico de malla direccional con diseño circular que muestra el enlace fuerte entre el módulo course y la acción view:

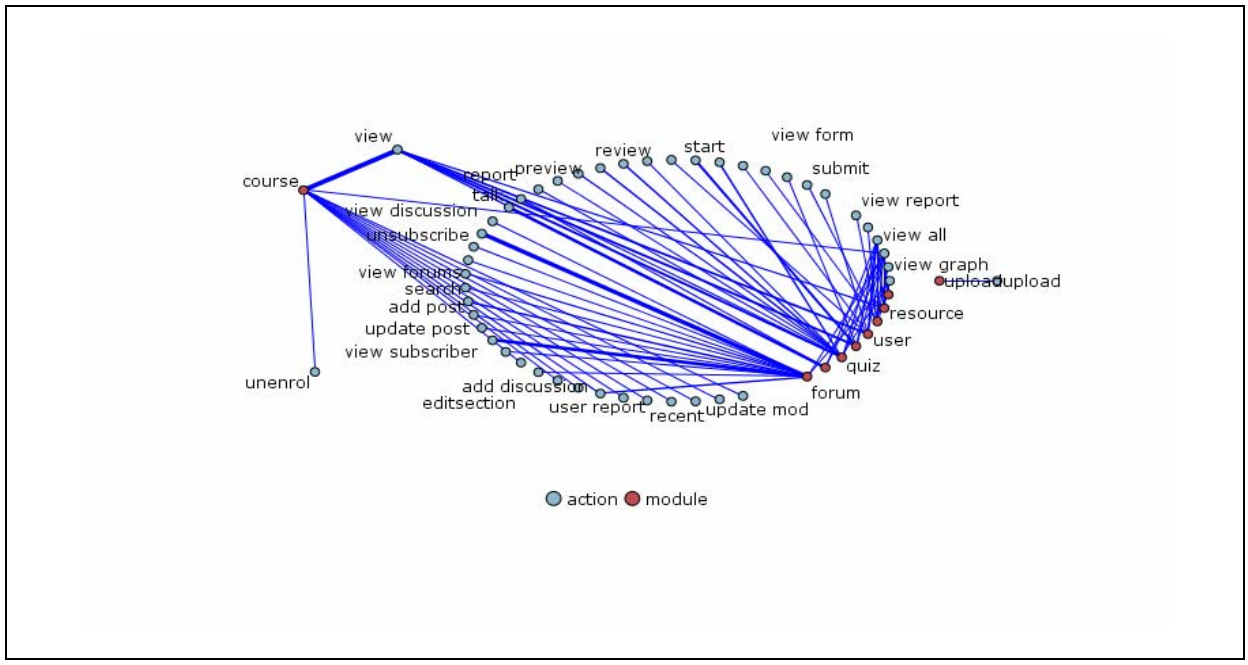


Figura 3. Enlace fuerte entre el módulo course y la acción view.

4. CONCLUSIONES DE LA APLICACIÓN DEL EDM SOBRE LA PLATAFORMA MOODLE DE EDUCACIÓN MUSICAL

A continuación y para finalizar el apartado destinado a la minería de datos educativos aplicados a nuestra plataforma de trabajo online en educación musical en el nivel de secundaria obligatoria, mostramos en una tabla el decálogo con los 10 puntos que sintetizan y justifican el empleo de estas técnicas estadísticas para la evaluación del modelo educativo orientado al trabajo con las TIC en el ámbito de la música en escuelas formales:

1. De mayor a menor frecuencia de uso de la plataforma los cursos son: tercero de la ESO, segundo de la ESO, primero de la ESO y cuarto de la ESO. Una de las posibles explicaciones que destacamos en estos resultados es que se demuestra que el nivel de actividad musical en la plataforma en mayor en nivel superior (de los cursos en los que es obligatoria la asignatura de música), lo cual, mediante la minería confirme algo, que parece lógico, que a mayor edad, mayor nivel de rendimiento. Sin embargo, queremos resaltar, que los resultados obtenidos en cuarto, aparecen justificados porque el perfil del alumnado, que elige la música, frecuenta resultados académicos más bajos (alumnado de letras y de diversificación curricular) frente al de ciencias, que escoge por regla general el alumnado con una mayor nivel de destrezas. Hecho que justifica el que en cuarto de ESO no se obtengan los mejores resultados. Con estos resultados detectamos anomalías en determinados grupos en cuanto al aprendizaje lo cual nos permite monitorizar los objetos de aprendizaje en los que su uso y rendimiento académico por parte del alumnado es deficiente.
2. El profesor-investigador de la asignatura desempeña aproximadamente una quinta parte de la actividad registrada en la plataforma (un 20% más o menos), quedando las otras cuatro quintas partes registradas por el alumnado. El tiempo de aplicación de la plataforma, aunque ha sido suficiente para nuestra investigación y los límites que la conforman por cuestiones de magnitud y espacio físico, debe continuar para aumentar la frecuencia de uso de la plataforma educativa-musical online por parte del alumnado. Un mayor uso, desde las casas incrementaría notablemente esta cierta desproporción entre la actividad del profesor en la plataforma y el conjunto del alumnado de primero a cuarto de ESO. En ese sentido y para potenciar el uso de nuestra herramienta informática, realizamos dos tutorías virtuales, una al final del segundo trimestre del curso y otra al final del tercero a través del chat de intercambio de la información.
3. Los módulos que presentan una mayor frecuencia de uso son, ordenados de mayor a menor: foros, lecciones musicales, exámenes musicales, chat, enlaces con información musical.

4. Las acciones que presentan una mayor frecuencia de uso son las de tipo exploratorio, como *ver foros* o *ver temas de debate de los foros*. En el caso de los chats, la acción mayoritaria es *hablar*. El motivo por el que destacan las acciones de tipo exploratorio, no sólo aparecen justificadas porque es la acción previa a cualquier otra en la plataforma sino porque vemos que la herramienta resulta novedosa y requiere de un mayor uso e implantación no sólo en nuestra área sino también en las demás: lengua, matemáticas, historia, plástica, educación física, para que el alumnado potencie su mejor y mayor uso dentro de las clases diarias que se imparten en los centros educativos y que pueden contar con una herramienta, como la plataforma educativa online para seguir trabajando, no sólo desde el centro sino también desde nuestra propia casa.
5. La herramienta del chat, requiere de la necesidad de un mejor enfoque para que su uso no sea un juego sino una potente herramienta de aprendizaje mixto o *blended-learning*, es decir, tanto presencial como a distancia.
6. Los modelos tipo clúster como el K-medias nos permiten conocer grupos de inactividad y aprendizaje deficiente, así como anomalías en la adquisición de conocimiento, por hacer categorías con el alumnado que no emplea módulos de especial importancia para el aprendizaje musical en red como son: el chat de intercambio de la información, el test de conocimientos musicales, los enlaces webs para la búsqueda de la información o las lecciones musicales con los contenidos programados en el curso.
7. Una de las limitaciones que ha tenido nuestro estudio en la fase de proceso es derivada de una falta de hábito por parte del profesorado de usar este tipo de herramientas de trabajo en red, dentro de la enseñanza secundaria obligatoria, lo cual, sumado al tiempo de entrenamiento del alumnado, para la adquisición de competencias básicas orientadas al aprendizaje del manejo de este tipo de soportes informáticos resta de manera considerada tiempo para el desarrollo de una actividad intensa y continuada en el tiempo.
8. Otra limitación importante es la deficiencia de los equipos informáticos y la imposibilidad de su mantenimiento por parte de una sola persona (el llamado coordinador TIC, el cual cuenta con tan sólo tres horas de reducción laboral a la semana), por lo que planteamos, como tarea organizativa e indispensable, la creación de grupos de trabajo (de al menos cuatro o cinco personas por centro), con carácter permanente y estable para mantener tal infraestructura, permitir su correcta utilización y funcionamiento orientados al desarrollo de buenas prácticas de aprendizaje digital en centros TIC y promover actividades en las diferentes áreas con las nuevas tecnologías y su implementación en plataformas telemáticas.
9. Planteamos como una opción posible de cara al futuro, la creación de centros TIC, pero con varias aulas específicas (tres o cuatro, en lugar de un número más numeroso y más difícil de mantener, por el grado de deterioro diario, al que está sometido el material informático), para su utilización por parte del profesorado y que cuenten con mayores posibilidades técnicas como un proyector digital y equipo de sonido de alta gama, unido a un buen sistema de microfonía, lo cual elimina la fatiga y el cansancio del aparato fonador del profesorado, el cual adolece frecuentemente de problemas de afonía y lesiones respiratorias por tener que hablar durante períodos prolongados de tiempo con una fuerte intensidad sonora.
10. El empleo de todas las técnicas, algoritmos y modelos estadísticos que empleamos en nuestro estudio validan la hipótesis de que el tipo de aprendizaje de alumnado está más basado en la exploración que en otro tipo de aprendizaje más útil de cara a la adquisición de conocimiento musical como la realización de trabajos colaborativos, las evaluaciones recíprocas y los test de conocimientos musicales. Esta averiguación en plataformas de teleformación musical se hace necesaria porque los algoritmos que utilizan los programas de la minería de datos que utilizamos (Statistica, Weka y Clementine) mejoran la eficiencia y el rendimiento de populares programas en el campo educativo y de las ciencias sociales como el paquete estadístico SPSS. El trabajo con grandes volúmenes de información, con numerosas variables y registros para analizar plantea el uso inevitable de otro tipo de software más adecuado, más rápido y flexible básicamente por su arquitectura, solidez y técnicas de muestro de la información.

Tabla 4. Decálogo con las conclusiones de la minería de datos educativos.

En resumen, planteamos la necesidad de la integración de los procedimientos estadísticos en minería de datos educativos dentro de las propias plataformas de teleformación, de forma que en cualquier momento dispongamos de la tecnología necesaria para detectar deficiencias el proceso de enseñanza-aprendizaje online, monitorizar objetos de aprendizaje, optimizar la configuración de la navegación en la herramienta y procurar que la adquisición de conocimiento útil en la asignatura mediante nuestro modelo educativo sea más fácil, sencillo y directo, sin necesidad de explorar la herramienta durante tanto tiempo para hacer un uso óptimo de la misma.

5. REFERENCIAS

Area Moreira, M. (2004). *Los medios y las tecnologías en la educación.* Madrid: Ediciones Pirámide.

De Pablos, Juan (1996). *Tecnología y Educación. Una aproximación sociocultural*. Barcelona: Editorial Cedecs.

Espigares Pinazo, M. J. y García Pérez, R. (2006). Educación musical con TICs en Escuelas Multiculturales. Almería: Actas del *I Jornadas Internacionales de Educación Intercultural* (VI Jornadas de Educación Intercultural) "convivencia y mediación intercultural".. http://www.ual.es/GruposInv/EducacionIntercultural/ (en prensa y CD-ROM). ISBN: 978.84.8240.826.2.

Proyecto colaborativo sobre EDM. Documento electrónico en: http://www.educationaldatamining.org/index.html (consultado 27/09/07)

Proyecto WEKA. Documento electrónico: www.cs.waikato.ac.nz/ml/weka (consultado 4/12/07).

Romero, C., Ventura, S. y García E. (2008). Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. Computers & Education. 51(**1**), 368-384.

Sierra Araujo, B. (2006). *Aprendizaje automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka.* Madrid: Editorial Pearson Prentice Hall.