Técnicas de Minería de Datos aplicadas al diseño de un Curso de Estadística

L.Lanzarini , J.Maulini, A.Villa Monte, L.Corbalán
Instituto de Investigación en Informática LIDI
Fac.de Informática – UNLP
{laural, jmaulini, avillamonte, corbalan}@lidi.info.unlp.edu.ar

M. Grossi
Depto. De Computación
Fac.de Ingeniería - UBA
{mdg7501@yahoo.com.ar}

Resumen

El presente trabajo utiliza estrategias pertenecientes al área de la Minería de Datos para analizar la metodología de enseñanza aplicada hasta el momento en un curso de Probabilidades y Estadística básico.
También se propone la utilización de una herramienta de software que reemplace la manera de realizar los trabajos prácticos. Con esto se espera poder contar con información permanentemente actualizada del desempeño de los alumnos y a la vez incorporar una metodología de trabajo que favorezca tanto a alumnos como docentes.
Los datos recolectados serán utilizados para obtener reglas de asociación que permitan mejorar el material y/o los procesos de aprendizaje.
Palabras claves: Minería de Datos Educativa, Enseñanza de Estadística.

1. Introducción

La Minería de Datos es una de las partes más importantes del proceso de extracción de conocimiento que permite obtener patrones útiles y novedosos, previamente desconocidos a partir de la información disponible [HER05]. Su aplicación en el ámbito educativo ha permitido caracterizar a los distintos actores que intervienen en los procesos de enseñanza-aprendizaje [ROM05].

Este trabajo muestra en particular la aplicación de determinadas técnicas de esta disciplina con el objetivo de evaluar la pertinencia y calidad del material desarrollado para un curso de Probabilidades y Estadística básico.

Además, se introduce una nueva propuesta metodológica centrada en una herramienta de software que facilite y registre la realización de actividades por parte de docentes y alumnos.

Por su intermedio se espera:
- Recolectar información referida a la actividad de los alumnos en el curso.
- Mantener la atención de los alumnos en la asignatura.
- Brindar una metodología para la resolución de los ejercicios.
- Facilitar la selección de ejercicios que forman la guía de trabajos prácticos.

Este artículo está organizado de la siguiente forma: en la sección 2 se describen las características del curso que ha sido objeto de estudio a lo largo de toda esta presentación junto con las conclusiones obtenidas a partir de estrategias de Minería de Datos; la sección 3 describe brevemente algunos trabajos relacionados con herramientas de software que podrían ser de interés en este tema; la sección 4 describe al cambio propuesto en la metodología de enseñanza; la sección 5 describe la herramienta de software a utilizar y finalmente la sección 6 resume las conclusiones.

2. Descripción del curso objeto de estudio

Este artículo propone una estrategia para la mejora de los procesos de enseñanza-aprendizaje de la asignatura Matemática 3 que se desarrolla en la Facultad de Informática de la Universidad Nacional de La Plata.

2.1. Características de la Asignatura

- Es una materia obligatoria de 2do. año para todas las carreras de la Facultad de Informática, excepto Ingeniería en Computación.

- Se trata de un curso de Probabilidades y Estadística de 70 hs. dictado de manera presencial.
- Su duración es de un cuatrimestre aunque también se ha dictado de manera intensiva.
- Tiene como correlativas previas a las asignaturas de matemáticas de primer año.
- Ninguna asignatura de 3er. año la tiene como correlativa previa.
- Constituye la última materia (de tres asignaturas cuatrimestrales), en la cual los alumnos reciben obligatoriamente conceptos de matemáticas.

2.2. Estrategia de Enseñanza

Los cursos de Matemática 3 analizados en el presente trabajo son los dictados durante 2009. En dicho año, la misma asignatura fue dictada tres veces: durante el mes de febrero en modalidad intensiva (4 semanas de 5 hs. diarias), durante el 1er. semestre en la Sede Tres Arroyos y durante el 2do. semestre para alumnos recursantes de La Plata.

A lo largo del año se introdujeron cambios en la modalidad de desarrollo de las clases.

En primer lugar y con el objetivo de incrementar la participación de los alumnos, se incorporaron herramientas basadas en WEB a través de la inserción del curso a la plataforma WebUNLP. Esto permitió contar con un entorno específico para los alumnos de Matemática 3 a través del cual, se comparte el material de clases y se definen las actividades a realizar.

El correo electrónico fue utilizado como herramienta de asesoría virtual. Por su intermedio, los alumnos pueden realizar consultas a cualquiera de los docentes.

Todo el material del curso se encuentra digitalizado; incluso las teorías han sido publicadas en dos formatos distintos para facilitar su acceso e impresión. Los ejercicios prácticos se encuentran totalmente resueltos en formato digital. La respuesta final (sólo el valor) de cada ejercicio se publica junto con el enunciado para permitir que el alumno realice

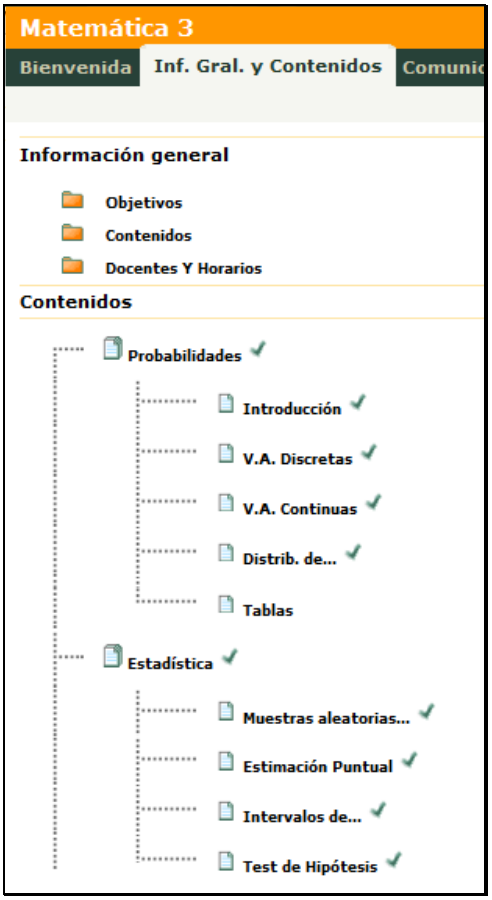

Figura 1. Información General y Contenidos del Curso de Matemática 3.

una primera auto-corrección. Una vez agotados los plazos de realización de la práctica, se le da difusión a la resolución completa de la mayoría de los ejercicios.

Para poder analizar el avance de los alumnos se implementaron dos estrategias con carácter obligatorio. La primera consiste en la entrega en papel de un número reducido de ejercicios; la segunda es la realización de autoevaluaciones luego de la finalización de cada tema.

Para facilitar la compresión por parte de los alumnos, los contenidos de la materia se encuentran divididos en dos partes: una primera parte correspondiente a Probabilidades y otra correspondiente a Estadística. La Figura 1 muestra, en cada caso, los temas principales involucrados. Cada sección incluye el material teórico y práctico correspondiente.

2.2. Características Principales de los alumnos involucrados

Si bien el curso está dirigido a alumnos de 2do. año, el hecho de que ninguna asignatura de 3er. año tenga como correlativa previa a Matemática 3 permite a los alumnos postergar su realización sin atrasarse en sus estudios.

Esto incrementa el número de alumnos recursantes dificultando la definición de horarios de clases sin superposición. En general se trata de personas que se inscriben en la asignatura y la abandonan sin completar las instancias de evaluación final.

2.3. Análisis de los cursos 2009

La información correspondiente a las actividades de los 240 alumnos de los cursos 2009 de Matemática 3 fue registrada a través de la plataforma WebUNLP.

Para cada alumno se consideró, además de las calificaciones obtenidas en las distintas actividades (ejercicios, autoevaluaciones y exámenes), la cantidad de recursos que descargó y la cantidad de días que demoró en realizar la descarga contando desde el momento de publicación del material.

El análisis de la información recolectada fue realizado con WEKA [Wit05] poniendo especial énfasis en reconocer las características del curso que mayor incidencia tienen en el resultado final.

El árbol de clasificación construido con el método J48, una variante del método C4.5 [QUI93][ZHU09], representado en la Figura 2, permite afirmar que la realización de las entregas de ejercicios durante el curso es una actividad importante en el proceso de aprendizaje (nodo raíz del árbol) ya que quienes las han aprobado han demostrado un buen resultado en el examen final. Para aquellos alumnos que obtuvieron calificación entre 4 y 6 en dichas entregas, el resultado de las autoevaluaciones define su desempeño.

También resulta de interés analizar el impacto del material del curso con respecto a las calificaciones finales.

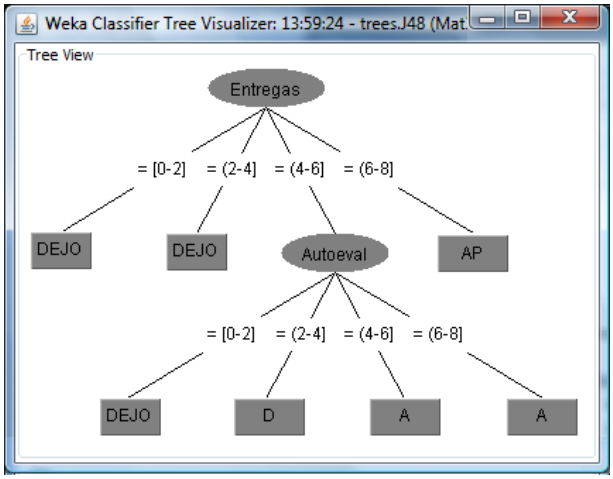

Figura 2: Arbol de Clasificación correspondiente a la nota final del curso.

En este caso se utilizó el algoritmo A Priori [AGR96] definido en WEKA con el fin de hallar reglas de asociación entre el nivel de descarga del material y la calificación final obtenida.

Los resultados muestran que las siguientes reglas

Si Descarga='MUY_ALTA'⇒Nota='APROB'

Si Nota='APROB'⇒ Descarga='MUY_ALTA'

tienen un Interés igual a 2.11, es decir que existe una fuerte dependencia positiva entre la condición de APROBADO y la descarga de contenidos MUY_ALTA (equivalente a más del 60% de los recursos). Lo mismo ocurre con los alumnos que desaprobaron o los que abandonaron el curso y una descarga de recursos reducida (inferior al 40%).

Lo anterior indica que el material del curso es adecuado para los alumnos.

Sin embargo, todas las mediciones referidas a las actividades son registradas con posterioridad a la realización de dichos trabajos no teniendo información alguna de resultados intermedios que permitan cuantificar el grado de aprendizaje adquirido.

Esto impide analizar el comportamiento de los alumnos que abandonan el curso sin completar las entregas prácticas. Este es un tema de sumo interés para los docentes ya que el porcentaje

de alumnos en estas condiciones es aproximadamente del 35%.

Además de la información recolectada a través de la plataforma WebUNLP, se encuestó a los alumnos una vez finalizado el curso y se les pidió que indicaran tres fortalezas y tres debilidades. Las fortalezas más destacadas se refieren a la utilidad del material para la compresión de los temas básicos así como la importancia de las actividades obligatorias (entregas y autoevaluaciones) a la hora de ganar experiencia en la resolución de problemas. Entre las debilidades del curso la más destacada es la falta de asistencia durante la realización de los ejercicios. Evidentemente, la relación actual de 1 docente de práctica cada 40 alumnos es insuficiente para responder todas las dudas que se presentan.

Como forma de enfrentar las dos situaciones mencionadas anteriormente: el análisis de las actividades realizadas por los alumnos durante el curso y la asistencia al momento de resolver los ejercicios prácticos, se propone el diseño y desarrollo de una herramienta de software que permita:

- Almacenar la información referida a las actividades de los alumnos durante la realización de los ejercicios.
- Brindar asistencia en la resolución de cada problema a través de consignas de menor complejidad.
- Ayudar en la adquisición de una metodología de trabajo basada en la identificación de los temas centrales de cada ejercicio que permitan al alumno dividir el problema a resolver en problemas más sencillos.

A continuación se describen algunas herramientas existentes y se discuten las dificultades de su aplicación. Luego, la sección 4 presenta la herramienta propuesta.

3. Trabajos Relacionados

En la actualidad, la cantidad de aplicaciones informáticas utilizadas en la enseñanza de temas de probabilidades y estadísticas es muy extensa.

Existen diversos proyectos que tienen como objetivo proveer recursos didácticos, interactivos y gratuitos en forma de applets. Por ejemplo, el Laboratorio Virtual en Probabilidades y Estadística [1] proporciona material explicativo en forma de lecciones audiovisuales de diversos experimentos relacionados con los distintos temas. Otro proyecto es Descartes[2] que permite crear lecciones interactivas instalables en modo local o accesibles desde un sitio web.

Si se buscan soluciones basadas en planillas de cálculo, Microsoft Excel y OpenOffice Calc permiten estudiar variables aleatorias, tanto discretas como continuas. Incluyen distintas operaciones que van desde el simple cálculo de probabilidades hasta la interpretación gráfica de cada distribución. [3, 4].

También existen diversas aplicaciones diseñadas en Javascript [5] para asistir a los usuarios en diversos experimentos numéricos que pueden resultar una herramienta muy útil para comprender técnicas y conceptos estadísticos específicos.

En cuanto a las aplicaciones de Software Libre, R [6] es una de las más conocidas por su potencia matemática y la diversidad de métodos estadísticos predefinidos que posee.

Por último, existen aplicaciones educativas de código cerrado (software propietario) con limitaciones de uso, modificación y/o distribución.

Si bien estas aplicaciones constituyen un importante complemento del material del curso, haciendo más ameno el aprendizaje de la asignatura, no permiten registrar las acciones realizadas por el alumno durante su uso. Sería de gran utilidad poder almacenar esta información para su posterior procesamiento.

El conocimiento por parte de los docentes de los pasos recorridos por los alumnos durante el proceso de enseñanza-aprendizaje es de suma

importancia para la toma de decisiones en la metodología a utilizar [GON04].

Lamentablemente, la recolección de información a partir de estas aplicaciones desarrolladas no es una tarea de menor complejidad. La reprogramación de las mismas, para lograr la obtención de dicha información, en la mayoría de los casos no es una tarea posible y en otros requiere tener un vasto conocimiento del lenguaje en el cual está desarrollada la aplicación.

A continuación se resume la herramienta de software propuesta para resolver este último aspecto.

4. Cambio propuesto en la metodología de enseñanza

Actualmente la práctica de la asignatura se realiza de la forma convencional, es decir, se publica un conjunto de enunciados que los alumnos deben resolver de manera presencial. Como control en el avance de cada tema, se exige la entrega de un número reducido de ellos.

Este enfoque obliga al alumno a asistir a clase a fin de poder consultar sus dudas y al docente a corregir y registrar las actividades en forma manual.

Por lo antes expuesto se propone modificar la manera de resolución de las actividades prácticas introduciendo el uso obligatorio de una herramienta informática que seleccione y presente los ejercicios correspondientes a cada tema. De esta forma, el alumno tiene la posibilidad de contar con un entorno de ejercitación práctica mucho más dinámico que el conjunto de ejercicios impresos.

Esto implica un gran desafío para la organización del curso ya que cada actividad o ejercicio debe ser clasificado en temas y subtemas identificando las dependencias entre ellos.

Para cada ejercicio se indicará el enunciado correspondiente y se realizarán distintas preguntas buscando guiar al alumno de manera gradual en la resolución del enunciado

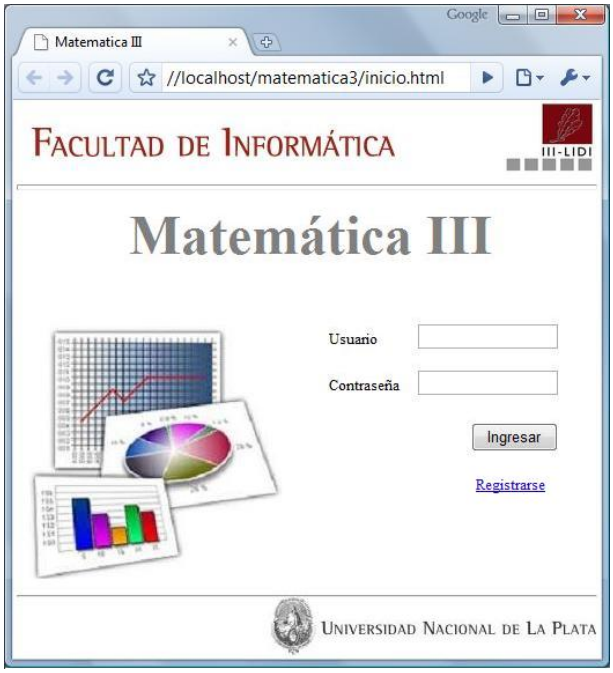

Figura 3: Pantalla de acceso al sistema.

completo. Con esto se espera introducir una metodología básica para la resolución de las actividades prácticas que será de utilidad no sólo para los alumnos sino también para los docentes.

El registro automático de las actividades realizadas por los alumnos permitirá a la cátedra disponer de información actualizada referida al desarrollo del curso pudiendo identificar

- El nivel de presencia y/o abandono de los alumnos.
- El grado de avance con respecto a las prácticas lo que tiene una fuerte relación con la comprensión de los temas vistos en la teoría.
- Dificultades generales en la resolución de ejercicios lo que puede llevar a rever los temas teóricos involucrados.

En lo que respecta a los alumnos, la herramienta reemplaza totalmente a las autoevaluaciones utilizadas hasta el momento ya que les permite conocer el resultado de sus actividades a medida que las van realizando.

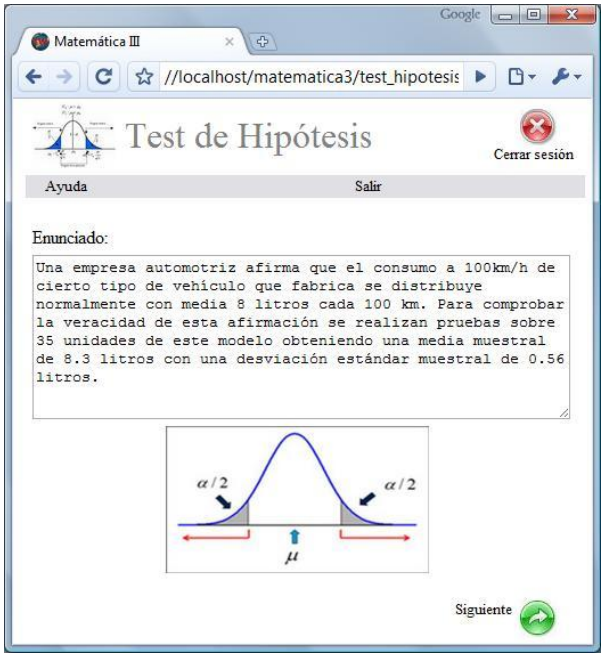

Figura 4: Presentación de actividad a realizar a través de la herramienta de software propuesta.

5. Herramienta de Software Propuesta

Se propone el desarrollo de una herramienta de software capaz de soportar la metodología presentada en la sección anterior.

El ingreso a la aplicación se realiza a través un nombre de usuario que permite separar docentes de alumnos (Figura 3).

En el caso de los docentes debe diseñar e ingresar el material práctico correspondiente clasificando cada ejercicio según el tema general y el subtema del cual se trate.

Cada ejercicio es presentado al alumno mediante un enunciado y opcionalmente una imagen asociada (ver figura 4)

La resolución presenta dos modalidades: ESTANDAR y DETALLADA. La primera es la utilizada por defecto y la segunda se aplica a aquellos alumnos con un alto índice de error en sus respuestas. En este último caso, se trata de preguntas más simples referidas al enunciado inicial que buscan hacer hincapié en los temas centrales con el objetivo de ayudar a construir una metodología básica de resolución.

Se busca que la corrección sea lo más automática posible por lo que, si bien es factible para el alumno responder a través de texto libre ingresado en un cuadro de edición, se recomienda el uso de preguntas de selección múltiple. En este último caso, para cada consigna debe indicarse la respuesta correcta. En caso de tratarse de valores numéricos y con el fin de brindar flexibilidad a la aplicación el docente debe indicar un rango de valores que serán utilizados para la generación aleatoria de opciones inválidas alternativas. La cantidad de alternativas para cada consigna es fija pero la posición en que aparece la respuesta correcta y los valores alternativos sugeridos por el sistema cambian su ubicación en la pantalla cada vez que se muestran.

Además, la complejidad de cada ejercicio se encuentra clasificada por el docente como NORMAL y AVANZADO a fin de presentarle al alumno los ejercicios más simples al inicio de su actividad.

Esto implica reconocer el grado de avance de los alumnos en el tema y por lo tanto son etiquetados inicialmente como INICIAL y en función de sus actividades durante el curso podrán adquirir la categoría de FORMADO. El cambio de categoría dependerá de la proporción de ejercicios de complejidad NORMAL que haya podido resolver.

También se registra la cantidad de intentos utilizados para cumplir con la consigna indicada.

El tiempo de resolución de cada ejercicio, si bien es registrado, no es considerado un factor decisivo del desempeño del alumno ya que, por tratarse de una aplicación con acceso remoto, en general el alumno opta por realizar algunos cálculos manualmente, fuera del sistema, para luego volver a conectarse e indicar la respuesta.

Mientras el alumno permanece conectado, el sistema automáticamente registra cada respuesta estableciendo la fecha y hora en que se realizó, la opción elegida y el resultado obtenido. Además, para cada ejercicio, se establece el momento en el que el alumno lo

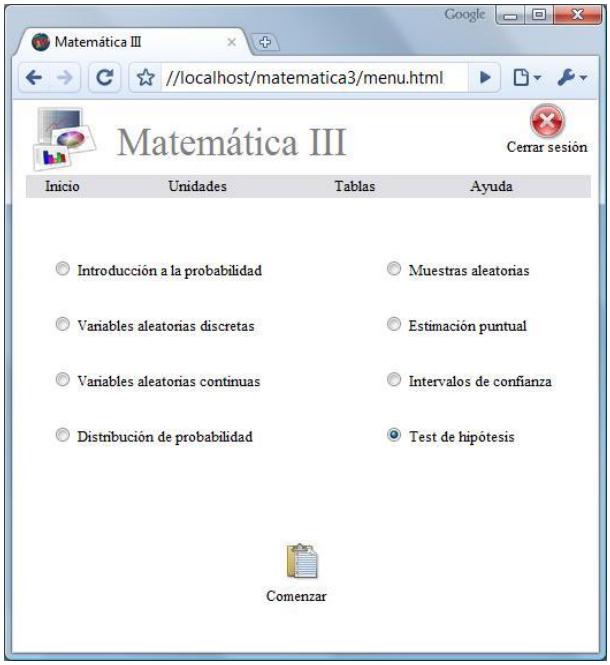

Figura 5: Pantalla para la elección del tema que se desea ejercitar (sólo disponible para los alumnos recursantes).

visualiza por primera vez y si fue finalizado o no.

Uno de los objetivos fundamentales de esta herramienta de software es brindar al alumno un entorno de ejercitación dinámico, capaz de mostrar ejercicios diferentes cada vez que se lo utiliza.

Para el caso de los alumnos recursantes es posible seleccionar el tema a ejercitar (ver figura 5). Para los alumnos que realizan la materia por primera vez, el recorrido de los temas es fijo.

La información recolectada podrá ser utilizada por el docente para monitorear las tareas de los alumnos así como para analizar el material presentado. En este último caso, a partir de la información disponible se espera obtener reglas de asociación que relacionen el nivel alcanzado por el alumno y la factibilidad de resolver un problema dado. Por ejemplo, una regla del tipo

Si NIVEL(1)='INICIAL' ⇒Ejercicio(4)=MAL

permitirá identificar que para los alumnos que en el tema 1 poseen nivel inicial es muy difícil resolver la actividad 4. Esto puede deberse a un error en el enunciado o un error en la clasificación del ejercicio.

Lo mismo ocurre con los alumnos de un determinado nivel que siempre aprueban un ejercicio, es decir, una regla del tipo

Si NIVEL(3)='FORMADO'⇒Ejercicio(15)=OK

puede deberse a la asignación de un nivel de dificultad AVANZADO cuando en realidad el ejercicio corresponde al nivel NORMAL.

También se considera factible obtener relaciones de dependencia entre temas y subtemas. Por ejemplo, una regla de la forma

Si TEMA(8)='APROB' ⇒Tema(1)=APROB

indicaría que la resolución de los ejercicios pertenecientes al tema 8 requieren de los conceptos vistos en el Tema 1.

6. Conclusiones

Se ha analizado el material y la metodología aplicada en tres cursos de Probabilidades y Estadísticas.

La aplicación de distintas técnicas de Minería de Datos a la información existente permite afirmar que el material utilizado es relevante para la aprobación del curso y que las actividades propuestas tanto en modalidad presencial como a través de la plataforma WebUNLP se encuentran estrechamente relacionadas con la calificación final del curso.

Se ha propuesto un cambio en la metodología de resolución de actividades prácticas reemplazando el conjunto de ejercicios convencional, indicados en papel, por una herramienta informática, actualmente en desarrollo, con capacidad para seleccionar las actividades a realizar por cada alumno de acuerdo a su avance en la materia.

Esta propuesta si bien presenta grandes ventajas requiere de esfuerzo adicional por parte de los docentes involucrados para clasificar y organizar el material existente.

Referencias

[AGR96] Rakesh Agrawal, Ramakrishman Srikant. Fast algorithms for Mining Association Rules. IBM Almaden Research Center. 1996.

[GON04] Carina Soledad González. Sistemas inteligentes en la educación: una revisión de las líneas de investigación y aplicación actuales. Revista Electrónica de Investigación y Evaluación Educativa RELIEVE, 10(1):3-22, 2004.

[QUI93] Ross Quinlan. C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, San Mateo, CA. 1993. ISBN: 978-1-55860-238-0

[ROM05] Cristóbal Romero Morales, Sebastián Ventura Soto, Cesar Hervás Martínez. Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.49-56. ISBN: 84-9732-449-8

[WIT05] Ian Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann. 2005. ISBN 0-12-088407-0

[ZHU09] Zhu Xiaoliang; Yan Hongcan; Wang Jian; Wu Shangzhuo. Research and application of the improved algorithm C4.5 on Decision tree. Test and Measurement, 2009. ICTM '09. International Conference on. Vol.2. pp184 - -187 ISBN: 978-1-4244-4699-5

[HER05] Henández Orallo, J.; Ramírez Quintana, M.J.; Ferri Ramírez, C. Introducción a la Mineria de Datos. España. Pearson Educación S.A.. 2005. ISBN 8420540919

[1]http://www.planetamatematico.com/index.php?option=com_content&task=view&id=118&Itemid=158

[2]http://platea.pntic.mec.es/jcarias/ccss2/02excel/10probabilidadexcel.htm

[3]http://home.ubalt.edu/ntsbarsh/stat-data/Javastatsl.htm

[4]http://www.r-project.org/

[5]http://www.freedownloadmanager.org/es/downloads/lognormal_gratis/