ISSN: 2448-6574

Evaluación de herramientas para predicción de deserción a nivel superior usando técnicas de minería de datos

Miguel Angel Couoh Novelo
miguelcouoh@hotmail.com

Edwin Miguel Poot Noh
edwinmiguel93@gmail.com

Raul Romualdo Cen Chan
raulcenchan@gmail.com

Resumen: Actualmente, los altos índices de reprobación y deserción reportados por las distintas universidades de nivel superior en México, indican graves problemas de retención del alumnado. Uno de los problemas que conduce al abandono escolar es el bajo rendimiento académico obtenido por los alumnos en las asignaturas; esta situación conduce a los alumnos a no tener oportunidades de aprobación en exámenes ordinarios y extraordinarios. Los cursos en línea no se encuentran exentos del problema de la deserción y es un hecho que matriculan más estudiantes, pero sufren de una deserción mayor que los cursos en la modalidad tradicional (Allen y Seaman, 2007) y (Neil Terry, 2001), al respecto Parker (2003) dice que la deserción en los cursos a distancia es de un 10 a 20 % mayor que en los cursos presenciales y es crítica desde una perspectiva tanto económica como de la calidad del programa. Aplicando una serie de encuestas a los alumnos y empleando algoritmos de minería de datos como el j48 se logró obtener indicadores que pueden predecir la situación de un grupo o de un alumno particular, permitiendo generar notificaciones o recomendaciones para aumentar la probabilidad de aprobación.

Palabras claves: Evaluación, Reprobación, Moodle, Weka, Minería de datos.

ISSN: 2448-6574

Planteamiento del problema

El problema que presenta la Unidad Multidisciplinaria Tizimín (UMT) de la Universidad Autónoma de Yucatán (UADY), es la deserción escolar, ya que la alta tasa de abandono de estudios influye en forma negativa en lo económico y en lo académico para la universidad.

De acuerdo al plan de estudios de la universidad, los alumnos tienen que cursar de manera obligatoria al menos 4 optativas durante toda la carrera. Las materias se apoyan obligatoriamente en un entorno de aprendizaje virtual VLEs (Virtual Learning Environments) en este caso Moodle, sin embargo, se ha observado que los alumnos no se sienten involucrados en los proyectos de aula, los debates del grupo, o simplemente suben o descargan tareas que tienen puntuación.

Los alumnos muestran muy poco interés en interactuar con la plataforma de apoyo Moodle, lo que implica que no estén al día con las actividades que el maestro realiza para complementar lo visto en clase. El problema obviamente no radica en que el alumno no tenga el dominio o la capacidad de usar la plataforma, ya que además de ser universitarios a todos se les da un curso de inducción al inicio de la carrera.

Los estudiantes no están siendo capaces de alcanzar el rendimiento esperado en cada semestre, lo que conduce a que el alumno, al reprobar una o más materias, tenga que recusarlas y de ese modo perjudica su avance curricular y cuando ya no puede cargar más materias, deserta de la UMT por lineamiento académico.

Justificación:

Moodle, al ser plataforma libre, permite añadir funcionalidades según la necesidad del usuario sin embargo, actualmente en la UMT no se ha aprovechado esta posibilidad, y la plataforma funciona de la manera tradicional, no logrando el interés de los alumnos y la finalidad de ser herramienta de apoyo.

ISSN: 2448-6574

En la actualidad instituciones han aprovechado los avances en el estudio de técnicas de minería de datos aplicadas a la educación (EDM), para contrarrestar la problemática de la deserción y han obtenido resultados favorables, como es el caso de las Universidades de la Sabana Colombia, Leonardo Da Vinci en Francia y Sidney en Australia, entre otros.

Una función de la herramienta integrada a Moodle será la predicción de notas y alertará al alumno y profesor del riesgo de reprobación, impulsando la idea de que el alumno tomará las medidas necesarias; sin embargo, también ofrece la función de retroalimentación en la que presentará los temas adecuados para reforzar el estudio conduciendo al alumno a organizar de manera correcta su esfuerzo.

Con la incorporación de la herramienta los alumnos tendrán la experiencia de un entorno diferente y útil, interactuando con una herramienta agente que les dé seguimiento como si se tratase de un tutor. Asimismo, el profesor sabrá qué alumnos están en riesgo de reprobación y qué temas no están siendo comprendidos, lo que generará mayor responsabilidad por parte de los alumnos y profesores.

Fundamentación teórica:

A continuación, se describe brevemente en que consiste cada una de las técnicas mencionadas.

- **Minería de datos:** La minería de datos es el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, datawarehouses, o cualquier otro medio de almacenamiento de información.
- **Arboles de decisión:** Es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

ISSN: 2448-6574

- **Modelos de regresión**: La representación de la relación entre dos (o más) variables a través de un modelo formal supone contar con una expresión lógico-matemática que, aparte de resumir cómo es esa relación, va a permitir realizar predicciones de los valores que tomará una de las dos variables (la que se asuma como variable de respuesta, dependiente, criterio o Y) a partir de los valores de la otra (la que se asuma como variable explicativa, independiente, predictora o X).
- **Clasificadores bayesianos**: En teoría de la probabilidad y minería de datos, un clasificador Bayesiano es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de ingenuo.
- **Sistema de gestión del conocimiento**: es un concepto aplicado en las organizaciones. Tiene el fin de transferir el conocimiento desde el lugar donde se genera hasta el lugar en dónde se va a emplear (Fuentes, 2010), e implica el desarrollo de las competencias necesarias al interior de las organizaciones para compartirlo y utilizarlo entre sus miembros, así como para valorarlo y asimilarlo si se encuentra en el exterior de éstas.
- **Aprendizaje máquina**: Es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

Objetivos:

Integrar una herramienta de predicción y retroalimentación a la plataforma e-learning Moodle de la UMT, con el fin de disminuir los índices de reprobación y/o deserción.

Objetivo Específico

1. Identificar el uso y aceptación de la plataforma Moodle por medio de encuestas aplicadas a estudiantes y profesores.

ISSN: 2448-6574

2. Conocer la arquitectura y funcionamiento de la plataforma virtual de la UMT.
3. Obtener el algoritmo de aprendizaje más eficiente para la minería de datos mediante la aplicación de validación cruzada.
4. Integrar el módulo o plugin de predicción y retroalimentación a la plataforma Moodle

Metodología:

1. **Recabar información de investigaciones hechas acerca de la deserción.**

Se recabó información de investigaciones hechas en la UMT sobre la deserción escolar, en la que los investigadores habían centrado su esfuerzo en detectar cuáles eran las materias que los alumnos reprobaban con mayor frecuencia y sus causas.

Se aplicaron tres encuestas de las cuales las dos primeras se dirigió a los profesores y alumnos, con la finalidad de conocer la cantidad de veces que los docentes utilizan Moodle para apoyarse en su enseñanza a los alumnos. Por otro lado, se aplicó la misma encuesta siguiendo la temática anterior para los alumnos. Con los resultados obtenidos de las encuestas aplicadas se busca conocer porqué los docentes y alumnos no utilizan con frecuencia la plataforma.

2. **Instalación de la plataforma Moodle y creación de cursos, cuestionarios y foros a modo de práctica.**

Se empezó por consultar información donde se explicaba la funcionalidad de las herramientas de Moodle. Para llevar la teoría a la práctica se instaló el LMS en un servidor web apache local, se crearon cursos, cuestionarios, foros etc. a modo de práctica.

Se realizaron funciones con el conjunto de herramientas del LMS entre ellas: la administración del entorno de aprendizaje (EA), de los participantes, la gestión de contenidos, la gestión del trabajo en grupos y la evaluación.

ISSN: 2448-6574

3. **Selección de las variables para la extracción de la información de la base de datos de Moodle.**

Se seleccionaron las variables que se utilizarán para extraer la información de la base de datos de Moodle, para medir el grado de interacción del alumno con la plataforma de las materias y alumnos de la Licenciatura en Ciencias de la Computación.

Posteriormente se programó un script en el lenguaje PHP que realice una consulta multitabla e integre los datos en una sola, después descargue la información en formato tipo CSV.

4. **Aplicación de algoritmos para generar un modelo de conocimiento.**

Se realizaron tareas de pre-procesado de datos, para transformar los datos originales a una forma más adecuada para ser usada por los algoritmos de Weka, en este caso las tareas consistieron en limpieza, transformación y discretización de los datos.

Después de haber realizado las tareas de pre-procesado, se obtuvo un archivo con el formato arff que contenía los datos de 255 alumnos con sus respectivas notas y número de interacciones con la plataforma en once cursos de la licenciatura en ciencias de la computación.
Debido a la gran cantidad de atributos recopilados (notas, interacciones), se realizó también un análisis para la selección de atributos y determinar cuáles son los que mayormente influyen en la variable de salida o clase a predecir (estado académico), para seleccionar estas variables se utilizaron los métodos disponibles en weka que son algoritmos de filtrado.

5. **Aplicación de Php-java-bridge para establecer un puenteo Java-Php**

Una vez que se eligió el modelo predictivo con base en arboles de decisión, se procedió a la construcción de una interfaz en Java Netbeans que permitiera interactuar con el modelo construido en Weka y a la vez comunicarse con un sistema web PHP (Moodle).

ISSN: 2448-6574

Se descargó el paquete binario phpjavabridge, se descomprimió el archivo Weka.war y se copió al directorio webapps de tomcat, se cargó el servidor desde cualquier navegador y corrió PHP y si este tiene una llamada a java se ejecutará en una página.

6. Módulo Moodle que se comunica con la aplicación Java Weka

Una vez que se logró la comunicación entre java y una página cualquiera php en la fase de pruebas, lo siguiente fue editar un módulo de Moodle; para ello se descargó la plantilla para plugins desde el sitio oficial de la plataforma y prácticamente se siguió el manual del desarrollador para la codificación ya que está estandarizado.

Para añadir el plugin y su funcionalidad a Moodle, se ingresó a la plataforma en modo administrador, luego se seleccionó la opción administración del sitio, luego plugins, y finalmente instalar complementos posteriormente se seleccionó el archivo zip y se agregó a la plataforma, al aparecer el mensaje de validación se seleccionó instalar complemento.

Resultados:

1. Resultados de los estudios realizados por investigadores de la UMT, y las encuestas a alumnos

Se aplicaron las encuestas a los alumnos de la carrera de Ciencias de la Computación, para identificar las materias más difíciles. Las causas, desde la perspectiva de ellos fueron: la confusión causada en la realización de ciertas actividades, la falta de bases anteriores para comprender los problemas, no entendían las explicaciones del profesor o no saber qué hacer cuando éste no se encuentra presente; también aceptaron que no le habían dedicado tiempo suficiente a los temas de la materia.

Entre los motivos clasificados en "otros", mencionaron hay demasiada información en las diapositivas y eso aburre, demasiada práctica, varios problemas por día, no hay mucho material

ISSN: 2448-6574

en internet para las tareas, no hay tiempo suficiente para acabar las actividades (Tareas y/o proyectos).

En otra encuesta aplicada a los mismos alumnos, los estudiantes dijeron que las materias con mayor dificultad son: Cálculo vectorial, teoría de computación, compiladores, y sistemas distribuidos, como se muestra en las tablas 7 y 8, que enlista todos los temas ofrecidos durante los semestres 1 y 2, los cuales se ordenaron de mayor a menor dificultad.

La columna "Opinan" corresponden al número de estudiantes que considera difícil la materia y la columna "Población" indican el número total de estudiantes que respondieron a la encuesta, la última columna muestra el "Porcentaje" de alumnos que opinaron respecto al total de la población.

2. Resultado, instalación de la plataforma Moodle y creación de cursos, cuestionarios y foros a modo de práctica.

Se corrió en un servidor local apache a la plataforma Moodle, presentado cursos creados como prueba, los cuales tuvieron actividades simuladas; también en el sistema se restauraron once cursos de semestres anteriores correspondientes a la LCC, los cuales ya se encontraban diseñados, trabajados e incluso guardaban todas las calificaciones de los alumnos.

3. Resultados de la validación cruzada

Se usó la validación cruzada (10 veces), la cual divide todo el conjunto de datos en 10 subconjuntos y aleatoriamente va tomando un conjunto para el testeo y el restante para entrenamiento. De esta forma, cada algoritmo se ejecutó 10 veces aprendiendo con un 90% de las instancias y se prueba con el 10% de instancias restantes.

ISSN: 2448-6574

4. Resultados de la interpretación del algoritmo j48

Se presentan las matrices de confusión generadas por este algoritmo respecto a la interacción de los alumnos de ciencias de la computación en cursos del semestre anterior en Moodle

=== Matriz de confusión ===

a b c ß clasificación

33 5 2 | a = MEDIO

4 31 5 | b = BAJO

1 6 31 | c= ALTO

Los índices en las filas y columnas de la matriz representan a las clases definidas en el modelado del usuario para el nivel general de interacción. Los elementos de la diagonal principal de la matriz representan el número de instancias clasificadas correctamente, es decir, Para el J48 con118 instancias (estudiantes), 33 tienen un nivel de interacción alto, 31 instancias nivel medio y 31 nivel bajo.

5. Resultados (aplicación de Php-java-bridge para establecer un puenteo Java-Php)

Como resultado de la programación en java, se obtuvo una aplicación capaz de comunicarse con Weka y aplicar cinco algoritmos al conjunto de datos y a la vez este permitió presentar la información en modo web. Para ello se utilizó la herramienta PHP Java Bridge con el objetivo de establecer la comunicación entre los dos lenguajes, a fin de presentar dicha información en la plataforma.

Conclusión:

A continuación, se resumen las principales conclusiones de las cuales se podrán reconocer con qué nivel fueron alcanzados los objetivos propuestos.

- Para conocer el problema de la deserción escolar se analizó previamente investigación de propuestas hechas por la UMT, información que fue de ayuda para conocer a lo que la

ISSN: 2448-6574

institución quería apegarse, surgieron otras ideas a partir de ahí para el desarrollo de la herramienta. Minería de datos fue el tema central que se abarco en la extracción de conocimiento.

- El objetivo era desarrollar e integrar un plugin a Moodle que fuera capaz de predecir notas y retroalimentar temas a los alumnos con la finalidad de aumentar las probabilidades de aprobación, anteriormente este objetivo se pretendía lograr con las cadenas de Markov, pero al revisar la literatura se definió a la minería de datos como la apropiada.
- Se encontraron trabajos relacionados al problema de la deserción escolar, y en general un campo mucho más extenso en el estudio de la minería de datos aplicados a la educación (EDM) que las cadenas de Markov, además de que esto significaba emplear algoritmos existentes. De igual manera se encontraron varios programas gratuitos y de paga para minería de datos, así como personas pioneras en el tema de deserción escolar.
- Para el proceso de integración, limpieza y discretizacion de los datos, se utilizó el software keel de licencia libre, pero no fue apto para la fase de minería de datos en Weka. Por otro lado, se comprobó que la interoperabilidad entre el lenguaje php y java es posible mediante un puente bridge y que a Moodle se le puede agregar funciones exclusivas de Java.
- Se logró la integración y la predicción, pero no la retroalimentación de temas en tiempo real, como se pretendía alcanzar.

Cabe mencionar que este trabajo fue resultado de un proyecto de investigación de intervención apoyado por el Verano Jaguar a alumnos del Instituto Tecnológico de Tizimín, realizado en la UMT de Tizimín.

Referencias:

Burke, 2000 R. Burke. Knowledge-based recommender systems. In J. E. Daily, A. Kent, and H. Lancour, editors, Encyclopedia of Library and Information Science, volume 69. Marcel Dekker. 2000.

ISSN: 2448-6574

Eduardo Adolfo Porcel, Gladys Noemí Dapozo y María Victoria López, (2010) "Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa".

Ernesto Pathros Ibarra García, Pablo Medina Mora, (2011), Creación de un modelo de predicción del desempeño académico de los alumnos de la facultad de ingeniería de la UNAM en el primer semestre.

González-Segura, C., Montañez-May, T., Chi-Pech, V., Miranda-Palma, C., & González-Segura, S. (2012). Analysis of Subjects with Greater Difficulty for University Students in the Area of Computer Science. *IJCSNS*, *12*(10), 62.

Hamalainen et al., (2004). Sistemas Tutores Inteligentes y reglas de asociación, y análisis de secuencias y regresión lineal en combination con reglas de red bayesiana de correlaciones. Int. Conf. on Intelligent. Tutoring Systems. Pp. 531-540.

Itmazi, J.A.S., (2005) “Sistema Flexible de gestión del e-learning para soportar el aprendizaje en las universidades tradicionales y abiertas”. PhD Thesis. University of Granada, Spain.

Merceron A., Yacef K. (2004), “Educational Data Mining: a case of study”, Universidad de Sydney, Australia.

Michael Pazzani, Daniel Billsus (1997) Learning and Revising User Profiles: The Identification of Interesting Web Sites.

Mladenic, D. (1999). Text-learning and related intelligent agents, in IEEE Experts, Special Issue on Applications of intelligent Information Retrieval, July-August, 1999.