

Clasificación de Indicadores de Interacción del uso de la plataforma Moodle para cursos de modalidad B-learning

Rejón Herrera, E.G. 1, Esparza Sánchez, R. 2, Pasos Ruiz, A. 3 Moreno Caballero, E.4
Universidad Autónoma de Yucatán, Facultad de Matemáticas
Tablaje Cat. 13615, Mérida, Yucatán, México.
1rherrera@uady.mx, 2rodrigo.esparza@uady.mx, 3apasos@uady.mx
Universidad Tecnológica Metropolitana, Circ. Colonias Sur 404, Mérida, Yucatán, México.
4eduardo.moreno@pa.utmetropolitana.edu.mx

Fecha de recepción: 30 junio 2015

Fecha de aceptación: 20 de agosto 2015

Resumen. En este trabajo se describe un caso de estudio de la clasificación de indicadores de interacción de los cursos de modalidad b-learning, que se imparten a nivel licenciatura en la Facultad de Matemáticas de la Universidad Autónoma de Yucatán. El objetivo es predecir el desempeño de un estudiante basándose en dichos indicadores. Utilizando el algoritmo de máquina de vector soporte se obtuvo un promedio de 84% de precisión. Además, se desarrolló una aplicación Java que predice el nivel de aprovechamiento académico de los estudiantes, a partir del modelo generado con el entrenamiento del algoritmo con cada conjunto de datos del caso de estudio. En la investigación realizada, se demuestra que es factible clasificar la frecuencia de uso de los recursos más utilizados por los alumnos y que tiene relación con su nivel de aprovechamiento académico.

Palabras Clave: B-learning, Minería de datos, aprendizaje supervisado, Máquinas de vector soporte, Weka.

Summary. In this work, a case study is presented for classification of interaction indicators for the b-learning courses taught at Math Faculty in Universidad Autónoma de Yucatan. The goal is predict student development based on those indicators. Using a support vector machine an accuracy of 84% was obtained. In addition, a Java application was developed to predict the student academic efficiency, using the generated model with the dataset of the study. In this research, it was proven that is possible to classify the frequency of the student's most used resources and that is has a direct relation with the academic efficiency.

Keywords: B-learning, Data Mining, Supervised Learning, Support Vector Machine (SVM), Weka.

1. Introducción

Una de las herramientas de software libre más utilizada en el ámbito académico, en la categoría e-learning, es la plataforma Moodle; la cual ha crecido en su uso y apoya la impartición de clases a nivel superior, tanto presenciales como en línea [1]. En dicha plataforma, los estudiantes pueden interactuar entre sí y con el profesor, con la ayuda de diversas herramientas y recursos, creando un entorno de aprendizaje de modalidad mixta o b-learning [2]. Sin embargo, se requiere conocer como los estudiantes interactúan con los recursos digitales disponibles, para realizar adecuaciones que mejoren su aprendizaje y establecer la relación de dicho comportamiento con su desempeño académico. También, se plantea, si un sistema en línea efectivo, puede proporcionar información útil para ayudar a los maestros a evitar que los estudiantes abandonen el curso o en su defecto, puedan utilizar patrones de acceso útil para mejorar el ritmo de la enseñanza [3]. En consecuencia, es necesario el desarrollo de herramientas y metodologías para analizar la interacción del alumno con el entorno virtual [4].

En este artículo, se propone utilizar las máquinas de vector soporte (Support Vector Machine, SVM), para clasificar los indicadores de interacción de los alumnos de cursos b-learning, que se imparten en la Facultad de Matemáticas de la Universidad Autónoma de Yucatán. El objetivo es demostrar que la frecuencia de uso de los recursos más utilizados por los alumnos en la plataforma Moodle, tiene relación con su nivel de aprovechamiento académico. Asimismo, su contribución es la implementación de un software para predecir el nivel de dominio de acreditación de los estudiantes [5], de las asignaturas que se incluyen en el caso de estudio. Lo anterior, a diferencia de otros proyectos, entre ellos, la implementación de técnicas de minería de datos para descubrir patrones de actividad en foros de discusión de plataformas e-learning [6] y, la evaluación de modelos

de clasificación de datos, de la formación basada en la web y las evaluaciones acumulativas de la enseñanza tradicional en el aula, para la predicción de su rendimiento en el examen final [7].

Las SVM representan una poderosa técnica empleada en la clasificación de datos y el análisis de regresión. Utilizan el aprendizaje supervisado, produciendo modelos matemáticos que son geoméricamente intuitivos y teóricamente fundamentados. Emplean un conjunto de pares entrada-salida, que adquieren una función de decisión que asocia a un nuevo dato, una etiqueta de clase dentro de las clases proporcionadas. Una ventaja notable radica en que obtienen un subconjunto de vectores de apoyo durante la fase de aprendizaje, que a menudo es sólo una pequeña parte del conjunto de datos original. Dicho conjunto de vectores es una tarea de clasificación dada y está formado por un conjunto compacto de datos [8]. Por lo tanto, el objetivo del aprendizaje supervisado es crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida, después de haber visto una serie de ejemplos, en este caso los indicadores de interacción.

Finalmente, el contenido del artículo está organizado de la siguiente manera: en la sección 2, se describe el marco teórico del proyecto; en la sección 3, se presenta la implementación y finalmente, en la sección 4, se exponen las conclusiones.

2. Marco Teórico

2.1. Minería de Datos

La minería de datos (Data Mining, DM) consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Asimismo, las herramientas tradicionales de minería de datos asumen que los datos que se usarán para construir los modelos contienen la información necesaria para lograr el propósito buscado de obtener suficiente conocimiento, que pueda ser aplicado al problema para su solución. Además, recientemente, ha incrementado el interés en utilizar sus técnicas en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales para comprender mejor a los estudiantes y el entorno en el que aprenden [9]. Hoy en día, uno de los mayores retos de las instituciones educativas es el crecimiento explosivo de datos sobre la educación y como utilizarlos para mejorar la calidad de las decisiones de gestión. Dichas técnicas de minería de datos son herramientas analíticas que se pueden utilizar para extraer conocimiento significativo, obtener información útil y proporcionar herramientas de análisis en los procesos de toma de decisiones [10].

De tal forma, DM se describe como un proceso de buscar patrones de manera automática [11], por lo que una de sus tareas es solucionar problemas de clasificación de información, utilizando algoritmos de aprendizaje automático, específicamente herramientas de aprendizaje supervisado.

2.2. Aprendizaje supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento, que consisten de pares de objetos (normalmente vectores), que constituyen los datos de entrada y los resultados deseados. La salida de la función puede ser un valor numérico (regresión) o una etiqueta de clase (clasificación). Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente [12].

Por lo tanto, busca aprender a partir de un conjunto de datos que se modelan como un conjunto de vectores de entrada-salida $\langle \mathbf{x}, y \rangle$ en donde \mathbf{x} es el vector de entrada que contiene varias características $\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle$. La salida se define dentro de un conjunto de valores predefinidos (y). En la figura 1, se puede observar gráficamente un conjunto de vectores entrada-salida que contienen dos características $\langle x_1, x_2 \rangle$ de entrada y una salida con dos posibles valores (cuadros amarillos, círculos rojos).

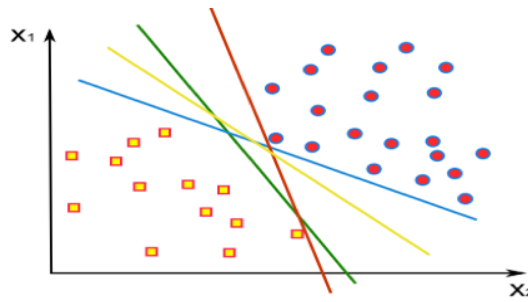


Fig. 5. Un clasificador lineal [13]

Uno de los modelos más sencillos utilizado en aprendizaje supervisado son las funciones lineales. Su objetivo es encontrar una función $f(x)$ que pueda discriminar los ejemplos en diferentes características. Por ejemplo, en la misma figura 1 se pueden observar varias funciones lineales que dividen del plano los ejemplos de la clase cuadros amarillos de los de la clase círculos rojos.

2.3. Máquinas de vector soporte

Las SVM son un conjunto de técnicas estadísticas que permiten clasificar una población en función de la partición en subespacios de múltiples variables. La idea principal es dividir de forma lineal un conjunto de múltiples dimensiones y son herramientas fundamentales en sistemas de aprendizaje automático, permitiendo el tratamiento de problemas actuales en reconocimiento de patrones y minería de datos tales como, reconocimiento y caracterización de texto manuscrito, detección ultrasónica de fallas en materiales, clasificación de imágenes médicas, sistemas biométricos, clasificación en bioinformática y en física de altas energías. Además, implementan reglas de decisión complejas, por medio de una función no lineal que permite mapear los puntos de entrenamiento a un espacio de mayor dimensión, por lo que en el nuevo espacio de características las clases son separadas por un hiperplano, siendo éste el que maximiza la distancia entre el mismo y los puntos de entrenamiento. Dicha distancia se denomina margen y los vectores son llamados de soporte.

Al existir varias funciones lineales, un problema importante es obtener la mejor función lineal. Esta es, la que maximiza la distancia entre el punto más cercano de una clase con el punto más cercano de la otra. Esta función se muestra en la figura 2 y se conoce como separador óptimo. Para encontrar el plano de separación óptimo se necesita resolver un problema de programación cuadrática (QP), que puede ser realizado por su problema dual introduciendo multiplicadores de Lagrange [14].

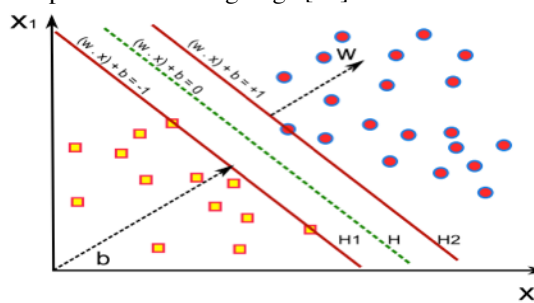


Fig. 6. Un clasificador óptimo [13]

En la Figura 2, se tiene la representación geométrica del problema de programación cuadrática mostrando H (separador óptimo) y los hiperplanos H_1 y H_2 . En el ejemplo, la cota de decisión es la línea que separa el espacio de entrada y está definido por la ecuación $w^T x_i + b = 0$. Donde w define el hiperplano de separación óptima, x_i es el vector de entrada que se asocia con una etiqueta (vector de salida) y b es el sesgo. La distancia entre el hiperplano de separación y el dato de entrenamiento más cercano, es el margen. La habilidad de generalización es maximizada, si el hiperplano de separación óptima, es seleccionado como el de separación. Por lo tanto, optimizar el margen geométrico significa minimizar la norma del vector de pesos. Al resolver el problema de programación cuadrática se trata de encontrar el hiperplano óptimo y dos hiperplanos (H_1 y H_2) paralelos. Las distancias entre H_1 y H , es maximizada y no existe ningún dato entre los dos. En el ejemplo, la distancia entre

H_2 es maximizada y algunos puntos de datos pueden estar sobre H y H_2 . Estos puntos de datos son los vectores soporte y participan de forma directa en definir dicho hiperplano de separación.

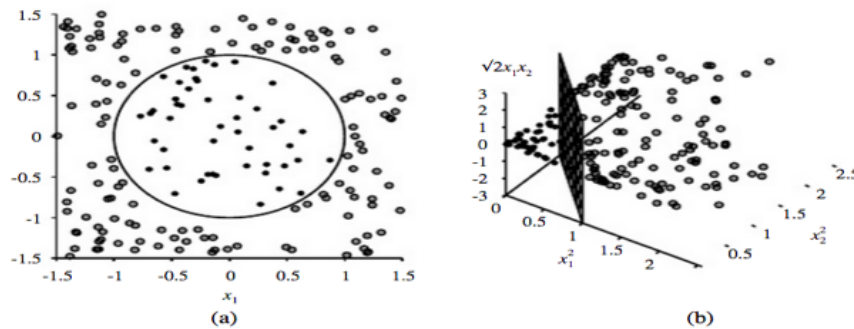


Fig. 7. a) Ejemplo de una función no lineal, b) Después de aplicar la función kernel [15].

A pesar de ser un modelo matemático con bases sólidas, los modelos lineales solamente funcionan si los datos son linealmente separables. Un ejemplo de una función no linealmente separable se puede observar en la figura 3a.

Por último, las SVM son una extensión de los modelos lineales para aproximar funciones no lineales. Esto se hace utilizando una o varias funciones kernel. Estas funciones incrementan la cantidad de características utilizando combinaciones de las ya existentes, permitiendo modelar funciones no lineales. Se ha demostrado que cualquier función no lineal con n características puede ser modelado como una función lineal de $n+k$ características, donde k es el número de funciones kernel [15]. Esto se puede observar gráficamente en la figura 3b.

3. Implementación

La Facultad de Matemáticas de la Universidad Autónoma de Yucatán cuenta con la plataforma e-learning Moodle, que es utilizada como herramienta de apoyo a las clases presenciales que se imparten en seis licenciaturas, tres posgrados y otros (estancias académicas, educación continua, etc.). Los recursos son habilitados por los profesores dependiendo del diseño de sus cursos y pueden incluir, contenidos y objetos de aprendizaje (course), materiales digitales (resource), asignación de tareas (assign), cuestionarios (quiz), foros de discusión (forum), vínculos a sitios de interés (links), wikis, correo electrónico (email), noticias (news), entre otros. Dicha plataforma cuenta con 314 tablas activas en su base de datos.

Se establece como caso de estudio, los cursos de licenciatura impartidos en la modalidad b-learning en los semestres de los años 2013 y 2014; seleccionados a partir de los siguientes criterios de inclusión: 1) mayor promedio de accesos por curso, 2) mayor cantidad de alumnos por curso y 3) mayor número de recursos habilitados por el profesor en cada curso. Por otro lado, se utilizan como indicadores de interacción, la frecuencia de acceso de los alumnos en cada uno de los recursos más usados (vector de atributos de entrada), así como el nivel de dominio obtenido para la acreditación de la asignatura (vector de salida).

A continuación, se describe la metodología utilizada en la solución del problema.

1. Extracción de datos.
2. Definición de criterios de inclusión.
3. Preparación del conjunto de datos.
4. Diseño y ejecución del algoritmo.
5. Elaboración del software predictivo del nivel de dominio.

3.1. Extracción de Datos

Consistió en obtener la información relevante a partir de las bases de datos existentes en la plataforma. Después de este proceso se determinaron los siguientes datos:

1. bitácoras de los registros de usuario por semestre.

2. total de accesos por curso y promedio de accesos por alumno/curso, así como el número de alumnos registrados en la asignatura.
3. número de recursos habilitados por el profesor en cada curso.
4. número de accesos por alumno/curso para cada recurso habilitado.
5. total de accesos a cada recurso utilizado en los cursos.

Tabla 1. Estadísticas de uso de la plataforma Moodle.

Descripción	2013		2014	
	Ene-May	Ago-Dic	Ene-May	Ago-Dic
Cursos licenciatura	81	78	78	84
Total de alumnos*	1,903	1,897	2,032	1,843
Total de Profesores *	62	72	57	66
Total accesos estudiantes	284,648	265,835	253,201	244,890

(*) Los alumnos pueden registrarse en más de un curso y los profesores pueden impartir más de un curso.

3.2. Definición de los criterios de inclusión

Se evaluaron los cursos de la Tabla 2, con los siguientes criterios de inclusión: 1) al menos, un promedio de cuatrocientos accesos por alumno/curso, 2) treinta o más alumnos registrados en el curso y 3) ocho o más recursos habilitados, de los cuales se seleccionan los más utilizados. En la tabla 2, se proporcionan las estadísticas del uso de la plataforma Moodle.

Tabla 2. Cursos con mejores estadísticas en los criterios de inclusión.

#	Cursos	Criterios de Inclusión		
		Promedio de accesos por alumno/curso (1)	Número de alumnos por curso (2)	Número de recursos habilitados (3)
Ene-May 2013				
1	APIS-01 (LCC)	584	32	9
2	AMAT- 01 (LIC)	491	13	14
3	APIS-02 (LIC)	480	31	9
4	AR-01 (LCC)	470	33	8
5	APIS-03 (LIC)	430	9	6
Ago-Dic 2013				
6	APIS-04 (LIS)	509	18	12

7	AES-01 (LIS)	382	33	9
8	ASB-01 (LCC)	366	15	10
9	Optativa -01 (LM)	348	7	7
10	Optativa -02 (LIC)	347	15	6
Ene-May 2014				
11	ASB-02 (LCC)	522	32	10
12	AAC-01 (LIC)	407	8	7
13	APIS-03 (LCC)	369	21	9
14	APIS-05 (LIS)	367	35	7
15	ASB-02 (LIS)	299	22	9
Ago-Dic 2014				
16	Optativa -03 (LIS)	436	5	10
17	ASB-01 (LCC)	393	15	11
18	APIS-04 (LIS)	350	37	10
19	ASB-02 (LIS)	321	19	8
20	Optativa -04 (LCC)	318	3	6

Debido a la confidencialidad de la información, sólo se menciona el área de conocimiento y la licenciatura al que pertenecen los cursos: APIS= área programación e ingeniería de software, AMAP = área de matemáticas, AR = área redes de computadora, AES = área entorno social, ASB= área software de base, AAC = área arquitectura de computadoras, LCC = licenciatura en ciencias de la computación, LIS = licenciatura en ingeniería de software, LIC = licenciatura en ingeniería en computación y LM = licenciatura en matemáticas.

Se establece el caso de estudio con las asignaturas: APIS-01 (LCC), APIS-02 (LIC), AR-01 (LCC) y ASB-02 (LCC). Lo anterior, porque dichos cursos propician el uso intensivo de la plataforma, habilitando diversos recursos para estimular el proceso de aprendizaje y cuentan con información suficiente para realizar el proceso de minería de datos (clasificación). A continuación, en la Tabla 3 se muestran dichas asignaturas.

Tabla 3. Asignaturas del caso de estudio.

#	Cursos	Criterios de Inclusión		
		Promedio de accesos por alumno/cursos (1)	Número de alumnos por curso (2)	Número de recursos habilitados (3)
1	APIS-01 (LCC)	584	32	9
2	APIS-02 (LIC)	480	31	9
3	AR-01 (LCC)	470	33	8
4	ASB-02 (LCC)	522	32	10

3.3. Preparación del conjunto de datos

Los datos de entrada se determinaron con las frecuencias de uso de los recursos más usados en las asignaturas del caso de estudio. Los indicadores de interacción corresponden a quiz, resource, assign y course, ya que representan el porcentaje más significativo del total de recursos habilitados. Dichos indicadores se relacionan con el nivel de dominio obtenido por cada alumno, para la acreditación de la asignatura. Dada la confidencialidad de la información, el Departamento de Control Escolar de la Facultad de Matemáticas, solamente proporcionó las calificaciones obtenidas por los alumnos los cuatro cursos que mejor cumplieron con los criterios mencionados. En el trabajo realizado, se eliminaron a los estudiantes que: a) no se enlistan en el acta de examen ordinario de la asignatura, b) los que aparecen sin derecho por no contar con un mínimo de 80% de asistencia en las clases presenciales y c) los que no utilizaron alguno de los recursos de mayor uso disponibles.

En la Tabla 4, se enlistan los niveles de dominio de acreditación de asignatura que establece el Modelo Educativo para la Formación Integral (MEFI) de la UADY.

Tabla 4. Niveles de dominio para la acreditación de una asignatura [5].

Puntaje	Categoría
90 - 100	Sobresaliente (SS)
80 - 89	Satisfactorio (SA)
70 - 79	Suficiente (S)
0 - 69	No acreditado (NA)

3.4. Diseño y ejecución del algoritmo

Posteriormente, se utilizó la SVM con la herramienta de minería de datos Weka (Waikato environment for knowledge analysis) [16] para la realización del experimento, incorporando el algoritmo LibSVM en Weka. Se utilizó validación cruzada para verificar la confiabilidad de la clasificación. En la Figura 4, se muestran un ejemplo del resultado de clasificación.

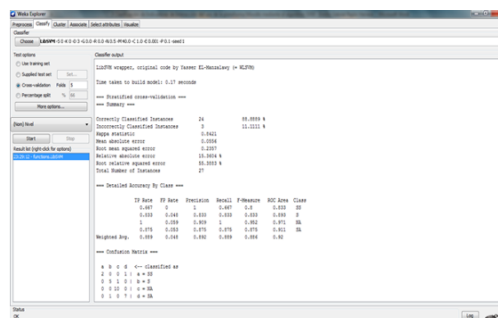


Fig. 4. Resultado de la clasificación del curso APIS-01 (LCC).

En la tabla 5 se pueden observar los resultados del experimento, donde se obtiene una precisión promedio del 84%.

Tabla 5. Resultados del caso de estudio.

#	Curso	Instancias correctamente clasificadas
1	APIS-01 (LCC)	88.8%
2	APIS-02 (LIC)	96.0%
3	AR-01 (LCC)	55.5%
4	ASB-02 (LCC)	96.6%

3.5. Elaboración del software predictivo de nivel de dominio

Utilizando Weka como librería externa, se elaboró un software en Java para generar el modelo predictivo de las asignaturas del caso de estudio. En la figura 5, se puede observar la ventana principal del Software Predictivo de Nivel Dominio de Acreditación.

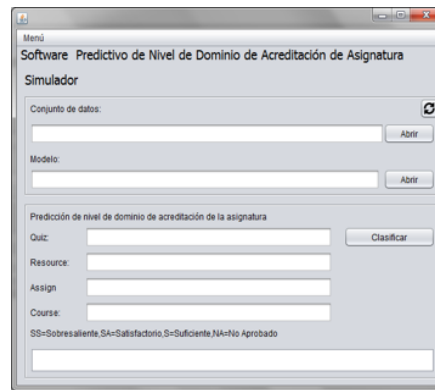


Fig. 5. Ventana principal del software predictivo elaborado.

4. Conclusiones y trabajos futuros

El resultado obtenido en la clasificación de las asignaturas del caso de estudio, es un promedio de 84% de precisión. Sin embargo, a pesar de cumplir con similares criterios de inclusión, el nivel de precisión del curso AR-01 (LCC) es sólo del 55.5%. Lo anterior, debido a que la frecuencia de uso, no siempre contiene información suficiente para predecir el aprovechamiento académico de los alumnos. Esto representa una desventaja del trabajo.

En la investigación realizada, se demuestra que es factible clasificar la frecuencia de uso de los recursos más utilizados por los alumnos en cursos de modalidad b-learning. De acuerdo con los resultados obtenidos, se pueden realizar recomendaciones a los estudiantes en el uso de los recursos digitales disponibles y a los profesores, en el diseño de las asignaturas para asistir a éstos en su aprendizaje.

Es conveniente señalar que para otros estudios, pueden aplicarse entrevistas al personal docente para conocer como incorporan el uso de la plataforma Moodle en sus cursos. Lo anterior, permitirá identificar los mejores indicadores para el proceso de clasificación; por ejemplo, se podrá determinar si las tareas asignadas son individuales o en equipo, de tal forma, dicho recurso podría ser relevante o no, como indicador de interacción para el estudio en cuestión. Además, también puedes aplicarse encuestas a los estudiantes para identificar sus estilos de aprendizaje.

Para trabajos futuros, se propone utilizar como indicador el tiempo de uso de los recursos disponibles, en vez de la frecuencia; así será posible realizar un análisis comparativo y determinar cual contiene mejores patrones de

clasificación. También, se pueden evaluar diversos algoritmos de clasificación supervisada, para determinar aquellos que generen los mejores resultados. Asimismo, dada las características de los cursos b-learning y su naturaleza interdisciplinaria, se pueden considerar líneas de investigación tales como estilos de aprendizaje y su relación con el diseño de contenidos de cursos en línea y la elaboración de algoritmos de clasificación supervisada, entre otros.

Referencias

1. Alonso Osorio, A.; Jiménez Builes J.: Uso de un sistema de administración del aprendizaje (LMS) libre como apoyo a los procesos de enseñanza y aprendizaje en instituciones públicas de educación superior. *Revista Avances en Sistema de Informática*. Vol. 6 No. 2 pp.5-10 (Septiembre 2009).
2. Olague Sánchez, J.; Torres Ovalle, S.; Morales Rodríguez, F.; Valdez Menchaca, A.; Silva Ávila, A.: Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales. *Revista Mexicana de Investigación Educativa*. Vol. 15, No. 45, pp. 391-421 (Abril-Julio 2010).
3. Hijón-Neira, R; Velázquez-Iturbide, J.; Barn, B.; Oussena S.: A Comparative Study on the Analysis of students Interactions in e-learning, *Eighth IEEE International Conference on Advanced Learning Technologies*. ICALT, pp. 20-22 (2008).
4. Falakmasir, M.H.; Moaven, S.; Abolhassani, H.; Habibi, J.: Business Intelligence in E-Learning (Case Study on the Iran University of Science and Technology DataSet), *IEEE, Software Engineering and Data Mining (SEDM), 2nd International Conference*. pp 473-477 (June 2010).
5. UADY.: Modelo Educativo para la Formación Integral, pp. 67, México. (2012).
6. Hsieh, Lu-shih.: Discovering Discussion Activity Flows in an On-line Forum Using Data Mining Techniques. *Tesis Doctoral, Universidad Nacional Sun Yan-Sen, China*. (2008).
7. Gamulin, J.; Gamulin, O.; Kermek, D.: Comparing classification models in the final exam performance prediction. *Invention and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention, IEEE. Croacia* (2014).
8. Cervantes Canales, Jair.: Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos. *Tesis Doctoral en Ciencias de Ingeniería Eléctrica, opción Computación. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México* (2009).
9. Jiménez, A.; Álvarez, H.: Minería de datos en la educación, Universidad Carlos III, *Revista ACM*, España (2010).
10. Kumar, V.; Chadha, A.: An Empirical Study of the Applications of Data Mining Techniques in Higher Education, *International Journal of Advanced Computer Science and Applications*, 2(3), pp. 80-84, (2011).
11. Witten, Ian; Frank, Eibe; Hall, Mark.: *Data mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems (2011).
12. Fernández, B.; Durán, E.: Buscador automático de material educativo en aulas virtuales, *IX Congreso sobre Tecnología en Educación & Educación en Tecnología*, p. 166-173, España, (2014).
13. Betancourt, Gustavo A.: Las Máquinas de Soporte Vectorial (SVM's). *Scientia et Technica Año XI*, No 27 (2005).
14. Cristianini, Nello; Shawe-Taylor, John.: *An introduction to support vector machines and other kernel based learning methods*. Cambridge University Press New York. (1999).
15. Russell, Stuart; Norvig, Peter.: *Artificial Intelligence, a modern approach*. Pearson Education (2003).
16. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten: The WEKA Data Mining Software: An Update; *SIGKDD Explorations*. Vol. 11, Issue 1, (2009).