Ciencias de la Información ISSN: 0864-4659 Vol. 46, No. 3, sept-dic, pp. 25 - 30, 2015

Aplicación de técnicas de descubrimiento de conocimientos en el proceso de caracterización estudiantil. *Application of knowledges discovery techniques in the student characterization process.*

Yoandry Pacheco Aguila
Yaima Fernández Segredo

El propósito de la presente investigación consiste en describir el proceso de descubrimiento de conocimientos sobre los resultados de las encuestas realizadas en el proceso de caracterización de una brigada estudiantil en la carrera Ingeniería en Ciencias Informáticas. Para el cual se aplicaron los pasos del algoritmo KDD, obteniéndose 5 subgrupos de estudiantes con características similares y bien determinadas entres sus miembros. Permitiendo obtener una visión más clara de los objetivos a definir en el proyecto educativo, así como realizar una adecuada planificación de las actividades en función de las metas específicas a alcanzar por cada subgrupo; con el fin de garantizar una adecuada interrelación entre los diferentes factores que intervienen en el éxito y la competitividad de las empresas que emplean a los egresados de la Universidad de las Ciencias Informáticas.

***Palabras clave**: descubrimiento de conocimientos en bases de datos, KDD, caracterización, proyecto educativo.*

RESUMEN

ABSTRACT

The purpose of this research is to describe the process of knowledges discovery applied to the results of the surveys accomplished in the characterization process of a student brigade of the career Engineering in Informatics Sciences. For which was applied the KDD algorithm steps, obtaining 5 subgroups of students with similar characteristics and well defined among its members. Allowing get a clearer view of the objectives to define in the educational project, as well as make proper planning of activities based on the specific goals to be achieved by each subgroup; in order to ensure proper interrelationship between the different factors involved in the success and competitiveness of enterprises that employ graduates the University Informatics Sciences.

***Keywords**: knowledges discovery in databases, KDD, characterization, educational project*

Introducción

Las empresas en constante desarrollo, con tendencias a extenderse, alcanzando una dimensión mundial que sobrepasan las fronteras nacionales, necesitan y exigen a la sociedad recursos humanos altamente calificados, tanto en el área profesional como en los aspectos éticos, morales y culturales, con el fin de garantizar una adecuada interrelación entre los diferentes factores que intervienen en el éxito y la competitividad de las empresas. Elementos que demandan de una adecuada proyección por parte de las universidades o entidades formadoras en el proceso de formación de cada individuo, a partir de los elementos que lo caracterizan.

Universidades cubanas tal como la Universidad de las Ciencias Informáticas (UCI) realizan el levantamiento y análisis de las características en función de los criterios individuales de cada estudiante, en el proceso de elaboración del proyecto educativo de cada brigada y del año en general. Estas actividades previas a la elaboración del proyecto educativo generan un gran volumen de información, del cual se tiene la oportunidad de descubrir conocimientos que servirán como soporte para la toma de decisiones en la elaboración del propio proyecto proyecto (OEI, 2007; OREALC/UNESCO, 2008). En tal sentido el propósito de este artículo es describir

el proceso de descubrimiento de conocimientos, aplicado sobre los resultados de las encuestas realizadas en el proceso de caracterización de una brigada estudiantil en su 1er año de la carrera.

Materiales y métodos

Se tomó como fuente de datos para la investigación los resultados de la encuesta aplicada, por la Universidad de las Ciencias Informáticas, en la recopilación de datos que caracterizan e influyen en el comportamiento y la actitud ante el estudio de los estudiantes que ingresaron al 1er año de la carrera.

Se aplicó el proceso de descubrimiento de conocimientos en bases de datos con apoyo de los algoritmos implementados en la herramienta Weka explorer (Molina y García, 2006; Martín et al., 2007; Witten et al., 2011), para conformar subgrupos de estudiantes con características semejantes y bien definidas entre sus miembros. A partir del cual se logró una proyección del trabajo diferenciado a realizar, por el colectivo pedagógico, en función de la formación integral de los estudiantes y acorde a sus necesidades formativas.

El proceso de descubrimiento de conocimientos en bases de datos (KDD del inglés Knowledge Discovery in Databases), es el proceso iterativo e interactivo, en el cual se usan, a través de medios automáticos y semiautomáticos, técnicas de aprendizaje inteligente sobre una o varias fuentes de datos (Molina y García, 2006), para extraer un cúmulo de información útil o patrones (Han y Kamber, 2006), que al ser interpretados y evaluados sobre un contexto determinado constituyen conocimiento que sirve como soporte a la toma de decisiones (Han y Kamber, 2006; citado en WebMining Consultores, 2011; Sánchez et al., 2011). Siendo las etapas previas a la extracción de la información decisivas en la fiabilidad e impacto del conocimiento extraído. (Reyes y García, 2005)

Cada etapa general del proceso KDD constituye un subproceso formado por una secuencia de etapas específicas en las cuales se aplicaron diferentes técnicas de datos implementadas en la herramienta Weka explorer. (Ver Figura 1)

Una técnica de datos constituye el enfoque conceptual para extraer información de los datos, que generalmente es implementada por varios algoritmos (Molina y García, 2006). En el proceso de KDD, estas técnicas se clasifican atendiendo a dos criterios:

• La fase del proceso de análisis de datos: Preprocesado o Minería de datos. (Molina y García, 2006)

• El objetivo del análisis de los datos: Supervisada o predictivas y No

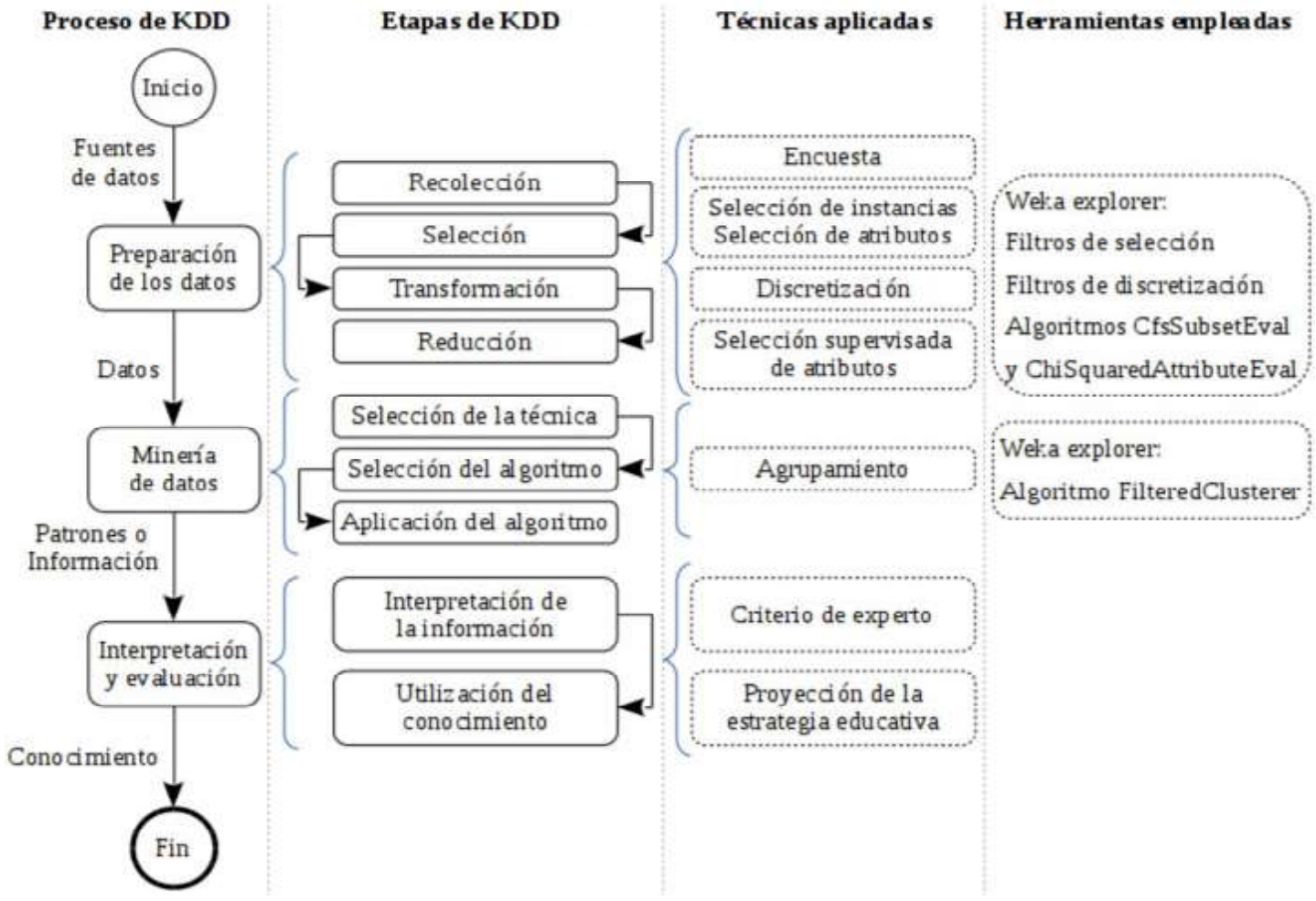

Figura 1: Proceso y etapas de KDD, técnicas y herramientas empleadas en la investigación.

supervisadas o descriptivas. (Weiss y Indurkhya, 1998 ; citado en Molina y García, 2006).

Definida la fuente de datos, el flujo y las etapas del proceso KDD, las técnicas de preprocesado y minería de datos aplicadas, así como las herramientas empleadas; se procede a la descripción de cada etapa del proceso y técnicas aplicadas en correspondencia con el objetivo del presente artículo.

Resultados y discusión

La Biblioteca Pública Municipal «Ñico López» del municipio La Lisa fue fundada en 1980 con cuatro salas de Aplicación de las técnicas de KDD en el proceso de caracterización estudiantil.

1ra etapa (Recolección de datos y dominio del problema):
Implica adquirir conocimiento del área de estudio del sistema y la meta a obtener. Se puede descomponer esta tarea en tres áreas:

• **Aprendizaje del tema:** El analista debe conocer el proceso detrás de la generación de la información para poder formular las preguntas correctas, seleccionar las variables relevantes a cada pregunta, interpretar los resultados y sugerir el curso de acción después de concluido el análisis.

• **Recolección de datos:** El analista debe conocer dónde se encuentran los datos correctos, cómo fueron obtenidos los datos de varias fuentes, cómo se pueden combinar estos datos y el grado de confianza de cada fuente.

• **Experiencia en análisis de datos:** El experto en minería de datos debe tener conocimientos adecuados en el uso de la estadística.

En esta etapa se aplicó la encuesta, como técnica de estudio observacional para recopilar datos por medio de un cuestionario previamente diseñado, compuesta por el conjunto de variables que caracterizan y de algún modo influyen en el comportamiento y la actitud del estudiante ante el estudio.

Listado de variables y su dominio de valores:

• Nombre del estudiante: (*texto*)
• Carné de identidad: (número de 11 dígitos)
• Grupo: (número de 4 dígitos)
• Dirección particular: (texto)
• Teléfono: (número)
• Valores morales que más admiran, sinceridad, honestidad, solidaridad, compañerismo, humildad, patriotismo, modestia, valentía, honradez, crítica y autocrítica: (si / no) por cada uno de los valores enunciados.
• Presencia de problemas personales que dificultan el estudio: (si / no)
• Presencia del apoyo familiar en mis estudios: (excelente / bueno / regular / no existe)
• Motivación por la informática: (muy motivado / motivado / poco motivado / no estoy motivado)
• Preferencia sobre la forma de estudio, por los textos, por notas de clases, por bibliografía orientada, por tele-clases, por el entorno virtual de aprendizaje: (si / no) por cada una de las formas enunciadas.
• Presencia de hábitos de estudio: (si / no)
• Prefiere estudiar: (solo / en grupo / en dúo)
• Frecuencia de estudio: (de forma sistemática / una semana antes de la evaluación / un día antes de la evaluación)
• Horas diarias dedicadas al estudio: (número)
• Me gusta cumplir las reglamentaciones y orientaciones: (estrictamente / parcialmente / hacer lo contrario)
• Me gusta participar en actividades de corte científico: (si / no)
• En clases suelo: (atender / distraerme con facilidad / dormirme / participar activamente)
• Interés por las informaciones nacionales e internacionales: (me mantengo informado / en ocasiones me informo / no estoy informado)
• Relaciones con los compañeros: (buenas / regular / malas)
• En el colectivo me gusta: (sobresalir / pasar desapercibido / relacionarme normalmente)

2da etapa (Selección y limpieza de datos):
Consiste en seleccionar un subconjunto de variables o datos de muestra, de los cuales se obtendrá conocimiento. Esta etapa se realiza con el fin de eliminar valores irrelevantes al objetivo, valores redundantes e inconsistencias en los datos de varias fuentes al juntarlos dentro de una sola base de datos.

En esta etapa se aplicó la técnica de selección de instancias para discriminar los datos correspondientes a los estudiantes que no pertenecen a la brigada objeto de análisis, quedando 27 instancias. Para ello se empleó el filtro de selección de instancias en Weka explorer, definiendo el atributo Grupo como criterio de selección.

También se aplicó la técnica de selección de atributos para eliminar campos tales como: nombre, carné de identidad, Grupo, dirección particular, teléfono y otros campos genéricos como las explicaciones y descripciones. Para ello se empleó el filtro de selección de atributos en Weka explorer, definiendo los atributos a discriminar sobre la base del criterio de experto, debido que a simple vista no constituyen información útil para el análisis y descubrimiento de conocimiento en el dominio del problema.

3ra etapa (Transformación de los datos):
Incluye operaciones básicas sobre los datos, como el filtrado para reducir ruido y decidir qué hacer con los datos faltantes. Otras tareas no tan evidentes están dadas por la redefinición de atributos, dado por la separación o compactación de estos.

En esta etapa se aplicó la técnica de discretización sobre el atributo «Horas diarias dedicadas al estudio», siendo definido como nuevo dominio de valores: menos de 2 horas / de 2 a 4 horas / más de 4 horas / ninguna. Para ello se se empleó el filtro de discretización en Weka explorer con los criterios establecidos para cada valor del nuevo dominio. También se transformó a mayúscula todos los valores de tipo texto.

4ta etapa (Reducción de datos y proyección):
En este paso el analista trata de buscar características más significativas para representar los datos en función de las metas del proceso y a su vez reducir las dimensiones de la base de datos.

En esta etapa se aplicó la técnica supervisada de selección de atributos para, de los atributos afines al objetivo del proceso, discriminar aquellos no significativos. Siendo eliminada del análisis la variable «RELACIONES CON LOS COMPAÑEROS», teniendo en cuenta el criterio de experto sobre los resultado obtenido de emplear los algoritmos «*CfsSubsetEval*» y «ChiSquaredAttribute Eval» disponibles en la herramienta Weka explorer (Martín et al., 2007; Witten et al., 2011) y que en todos los cuestionarios coincidió con el valor «BUENA».

5ta etapa (Elegir la técnica de minería de datos): Consiste en definir la técnica de minería de dato a emplear en función del objetivo del proceso.
Usualmente son empleadas las siguientes técnicas: (Pautsch, 2009; citado en Orellana y Ochoa, 2012)

• **Síntesis:** Dado una gran cantidad de atributos, es necesario sintetizar los datos usando varias reglas características que simplificarán la construcción del modelo.

• **Asociación:** Los algoritmos en esta clase generan reglas que asocian patrones de transacciones con cierta probabilidad.

• **Agrupamiento:** Agrupar objetos dentro de clases a partir de sus características, maximizando la semejanza dentro de la misma clase y minimizando la semejanza entre clases diferentes.

• **Clasificación y predicción:** Categorizar los datos basándose en un conjunto de datos de entrenamiento y elaborar un modelo para cada clase. Este modelo sirve para clasificar los nuevos datos agregados a la base de datos.

Con el fin de lograr una mejor planificación del trabajo diferenciado, en el presente estudio se eligió la técnica no supervisada de agrupamiento para determinar subgrupos de estudiantes en los cuales se maximicen las semejanzas en un mismo subgrupo y las diferencias entre subgrupos diferentes.

6ta etapa (Elegir el algoritmo de minería de datos):
La tarea consiste en seleccionar el método a ser usado para la búsqueda de patrones en los datos. Esto refina el alcance de la tarea anterior para utilizar el algoritmo más adecuado que ayude a alcanzar el objetivo final.

En esta etapa se eligieron los algoritmos «FilteredClusterer» y «FarthestFirst» disponibles en la herramienta Weka explorer (Witten et al., 2011, cap.3-4). Ambos son usados como técnicas de agrupamiento para encontrar grupos homogéneos, permitiendo detectar los

Tabla 1. Resultados de la aplicación del algoritmo weka.clusterers. FilteredClusterer.

Variables	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
V. Sinceridad	si	si	si	si	si
V. Honestidad	si	si	si	no	si
V. Solidaridad	si	si	si	no	si
V. Compañerismo	si	si	si	no	si
V. Humildad	si	si	si	no	si
V. Patriotismo	si	si	si	no	no
V. Modestia	si	si	si	no	si
V. Valentía	si	si	si	no	si
V. Honradez	si	si	si	no	si
V. Crítica y autocrítica	no	si	si	no	no
Problemas personales	no	no	no	no	no
Apoyo familiar en mis estudios	excelente	excelente	excelente	excelente	excelente
Motivación por la informática	**muy motivado**	**muy motivado**	**muy motivado**	**muy motivado**	**muy motivado**
F.E. Por los textos	si	no	si	si	si
F.E. Por notas de clase	si	no	si	si	si
F.E. Bibliografía orientada	si	si	no	no	si
F.E. Tele-clase:	no	no	no	no	no
F.E. Por el entorno virtual	si	no	si	no	si
Hábitos de estudio	si	no	si	si	si
Prefiere estudiar	en grupo	en grupo	en dúo	en dúo	solo
Frecuencia de estudio	sistemáticamente	sistemáticamente	sistemáticamente	sistemáticamente	una semana antes del examen
Horas diarias dedicadas al estudio	mas de 4	de 2 a 4	de 2 a 4	de 2 a 4	de 2 a 4
Cumplir reglamentaciones y orientaciones:	estrictamente	estrictamente	estrictamente	estrictamente	estrictamente
Actividades de corte científico:	si	si	si	si	si
En clases suelo:	atender	atender	atender	atender	atender
Informaciones nacionales e internacionales:	buena	buena	regular	buena	regular
Relaciones con los compañeros:	buena	buena	buena	buena	buena
En el colectivo me gusta:	relacionarme normalmente	relacionarme normalmente	relacionarme normalmente	relacionarme normalmente	pasar desapercibido

estilos o preferencias predominantes en cada subgrupo.

7ma etapa (Minería de datos):
Aplicación del algoritmo seleccionado.
En esta etapa se aplicaron los algoritmos «*FilteredClusterer*» y «*FarthestFirst*» disponibles en la herramienta Weka explorer, sin embargo, a partir de la interpretación de los resultados arrojado por ambos algoritmos, se determinó como algoritmo adecuado para el caso de estudio el *FilteredClusterer* para 5 subgrupos y el resultado se refleja en la tabla 1.

8va etapa (Interpretación):
Interpretación de los resultados obtenidos en la ejecución del algoritmo seleccionado.

El 1er subgrupo con 10 estudiantes para un 38% se caracteriza por:

- Alta afinidad por los valores éticos morales a excepción de la crítica y la autocrítica.
- Ser sistemáticos en los estudios, preferentemente en grupos, dedicándole más de 4 horas diarias y utilizando todos los medios disponibles a excepción de las tele-clases.
- Mantenerse informados sobre el ámbito nacional e internacional.

El 2do subgrupo con 2 estudiantes para un 8% se caracteriza por:

- Alta afinidad por los valores éticos morales.
- Ser sistemáticos en los estudios, preferentemente en grupos, dedicándole de 2 a 4 horas diarias y prefieren estudiar a través de la bibliografía orientada.
- Mantenerse informados sobre el ámbito nacional e internacional.

El 3er subgrupo con 5 estudiantes para un 19% se caracteriza por:

- Alta afinidad por los valores éticos y morales.
- Ser sistemáticos en los estudios, preferentemente en dúos, dedicándole de 2 a 4 horas diarias y no gustan de estudiar por bibliografías ni por tele-clases.
- Tener poco interés sobre aspectos del ámbito nacional e internacional.
- Pasar desapercibidos por la brigada.

El 4to subgrupo con 7 estudiantes para un 27% se caracteriza por:

- Valorar la sinceridad por encima de todos los demás valores éticos y morales.
- Ser sistemáticos en los estudios, preferentemente en dúos, dedicándole de 2 a 4 horas diarias y prefieren estudiar por las notas de clase y otros textos.
- Mantenerse informados sobre el ámbito nacional e internacional.

El 5to subgrupo con 2 estudiantes para un 8% se caracteriza por:

- Valorar en menor medida el patriotismo, la crítica y la autocrítica respecto al resto de los valores éticos y morales.
- Ser finalistas en los estudios y estudiar preferentemente solos, dedicándole de 2 a 4 horas diarias, utilizando todos los medios disponibles a excepción de las tele-clases.
- Tener poco interés sobre aspectos del ámbito nacional e internacional.

9na etapa (Utilización del conocimiento obtenido):
La aplicación de los patrones extraídos puede implicar uno de los siguientes objetivos:

- **Descripción**: La meta es simplemente obtener una descripción del sistema bajo estudio.
- **Predicción**: Las relaciones obtenidas son usadas para realizar predicciones de situaciones fuera de la base de datos.
- **Intervención**: Los resultados pueden conducir a una intervención activa en el sistema modelado.

Los resultados obtenidos en este estudio permiten definir, orientar y dirigir los objetivos y tareas a cada subgrupo en particular, con el fin de erradicar las deficiencias o aspectos negativos.

- Para el 1er subgrupo deben estar encaminadas al mejoramiento de la crítica y la autocrítica.
- Para el 2do subgrupo deben estar encaminadas al desarrollo de habilidades en otras formas de estudio que no sea solo la bibliografía orientada en clase.
- Para el 3er subgrupo deben estar encaminadas a la socialización en la brigada, el trabajo en equipo, uso de la bibliografía y el interés por la información de carácter nacional e internacional.
- Para el 4to subgrupo deben estar encaminadas hacia el trabajo en equipo, así como la formación y valoración de otros valores éticos y morales.
- Para el 5to subgrupo deben estar encaminadas hacia las actitudes patrióticas, críticas y autocríticas, sistematización el estudio e interés por otros aspectos nacionales e internacionales.
- En sentido general se evidencia que es necesario hacer un análisis de las tele-clases, dada la baja preferencia que le dan los estudiantes como material de estudio.

Conclusiones

A partir de la aplicación de las técnicas de preprocesado y minerías de datos; selección de instancias, selección de atributos, discretización y agrupamiento; en el proceso de descubrimiento de conocimientos en bases de datos para la caracterización estudiantil, se obtuvo 5 subgrupos de estudiantes con similares características y bien determinadas entres sus miembros, lo cual brinda al colectivo pedagógico de la brigada una visión más clara de los objetivos a definir en el proyecto educativo, así como realizar una adecuada planificación de las actividades, en función de las metas específicas a alcanzar por cada subgrupo.

Recomendaciones

Se recomienda realizar el mismo proceso al cierre del curso o en el desarrollo del proyecto educativo del próximo curso, si se mantienen aproximadamente los mismos miembros de la brigada. De este modo se podría comprobar la efectividad

o no de las actividades planificadas, en función de los objetivos propuestos a partir del conocimiento extraído en la anterior iteración.

Bibliografía

Han, J. y Kamber, M. (2006). Data mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. 2.a Ed. USA: Morgan Kaufmann. — ISBN: 1-55860-901-6

Martín, R., Ramos, R.M., Grau, R. y García, M.M. (2007). Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños. [en línea]. Vol. 9, No. 1. Disponible en: http://bvs.sld.cu/revistas/gme/pub/vol.9.(1)_01/p1.html. [Accedido 1 diciembre 2012].

Molina, J.M. y García, J.: (2006). Técnicas de análisis de datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka. España: Universidad Carlos III de Madrid.

OEI (2007). Proyecto educativo regional Cajamarca 2011-2021. 1.a Ed. Perú. —ISBN: 978-9972-2959-1-1

OREALC/UNESCO (2008). Situación Educativa de América Latina y el Caribe: garantizando la Educación de Calidad para Todos. 1.a Ed. Santiago de Chile: Salesianos Impresores S.A.

Orellana, A. y Ochoa, A.J. (2012). Vista de análisis usando la técnica de agrupamiento para el sistema integral para la atención primaria de salud. En: diciembre 2012, La Habana, Cuba. —ISBN: 978-959-212-811-8

Pautsch, J.G.A. (2009). Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. Grado. [en línea]. Argentina: Universidad Nacional de Misiones. Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/TFAGermanPAUTSCHFinal.pdf.

Reyes, J.F. y García, R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. [en línea]. Vol. VIII, No. 26. Disponible en: http://ingenierias.uanl.mx/26/pdfs/26_el_proceso.pdf. [Accedido 12 marzo 2013].

Sánchez, E., Alpuín, H., Ochoa, H.J. y Pozos, P. (2011). SDCA: System to detect cancerous abnormalities. [en línea]. Vol. 804. Disponible en: http://ceur-ws.org/Vol-804/11_LANMR11.pdf.

WebMining Consultores (2011). KDD: Proceso de Extracción de conocimiento. [en línea]. 10 enero 2011. Disponible en: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/. [Accedido 8 febrero 2013].

Weiss, S.M y Indurkhya, N. (1998) Predictive Data Mining: a practical guide. San Francisco: Morgan Kaufmann. —ISBN: 1-55860-403-0

Witten, I.H., Frank, E. y Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. 3.a Ed. USA: Morgan Kaufmann. — ISBN: 978-0-12-374856-0

Recibido: 12 de octubre de 2015.
Aprobado en su forma definitiva:
15 de diciembre de 2015

Yoandry Pacheco Aguila
Universidad de las Ciencias Informáticas.
La Habana. Cuba
Correo-e.: andypa@uci.cu

Yaima Fernández Segredo
Universidad de las Ciencias Informáticas.
La Habana. Cuba
Correo-e.: yfsegredo@uci.cu