

Análisis de la deserción estudiantil en la USB, facultad Ingeniería de Sistemas, con técnicas de minería de datos

G. Romero P¹ y A. Paredes Niz²
g.romero@unisimon.edu.co¹, a.paredes@unisimon.edu.co²

Resumen-En este trabajo se pretende abordar la investigación de diferentes técnicas de minería de datos para la solución de un problema en el entorno de la Academia con respecto a la deserción de estudiantes. Para la selección de la técnica de minería de datos se debe tener en cuenta una serie de criterios de acuerdo al entorno del problema, además de ello, hay diferentes algoritmos que implementan la técnica por lo cual es necesario investigarlos y seleccionar el que mejor se adecue a los requerimientos y necesidades desde la perspectiva del problema que se quiere solucionar

Palabras claves—Minería de datos, clasificación, agrupamiento, almacenes de datos, descubrimiento de conocimiento, deserción universitaria, perfiles de alumnos.

Abstract- In this work it is tried to embroider the investigation of different techniques from data mining for solution of a problem in the surroundings of the e-commerce of type B2C. For the selection of the technique of data mining are due to consider a series of criteria according to the surroundings of the problem, in addition to it there are different algorithms that implement a technique of data mining, thus is necessary to investigate them and to select the one that adapts better to the requirements and necessities from the perspective of the problem that need to be solved.

Keywords—Data Mining, KDD, Data Mining Techniques

I. INTRODUCCIÓN

Para obtener conocimiento es necesario partir de la materia prima, que son los datos, los cuales se encuentran disponibles en gran cantidad gracias a las tecnologías de información y las comunicaciones. Estos datos por lo general se encuentran en forma no refinada y para poder analizarlos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos. Para realizar un análisis en profundidad de forma automática, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos, sin embargo, es la transformación de los datos en conocimiento y la aplicación de este lo que genera valor para una organización. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

Para las organizaciones que realizan algún tipo de mercado, el conocimiento es algo imprescindible para tener éxito, por tanto encontrar asociaciones o correlaciones interesantes en los registros de las transacciones de negocios puede ayudar a la toma de decisiones en los proyecto de marketing, por lo cual invierten recursos técnicos y económicos para la

construcción o compra de herramientas para el análisis de información. Es en este entorno la minería de datos ofrece la posibilidad de llevar a cabo un proceso de descubrimiento de información automático.

Existen algunas herramientas que cubren aspectos generales del proceso de minería de datos y por lo general están orientadas a propósitos académicos presentando algunas dificultades en su adaptación tales como, modificaciones en el código fuente, tipo de lenguajes implementados, tipo de licencia de uso, formato de entrada de datos, entre otros, lo que dificulta su uso simple por parte de las diferentes organizaciones.

Para brindar una solución acorde a las necesidades de las empresas es necesarios entender los objetivos y requerimientos desde la perspectiva de lo que se busca, convirtiendo entonces este conocimiento en la definición de un problema de minería de datos, ya que dependiendo del problema de información que se desea solucionar, existe una serie de técnicas que son aplicadas en la solución de diversos problemas.

Para la escogencia de una técnica de minería de datos se debe tener en cuenta una serie de consideraciones tales como el entorno del negocio, las fuentes de los, tipos de clientes, entre otros, que afectan el desempeño de estas. Además de ello existen diferentes algoritmos que implementan las técnicas de minería de datos y cuya utilización afecta el rendimiento de estas, por lo cual es necesario estudiarlos para determinar cual es le mas indicado

En lo que resta del documento se hablará de proceso de descubrimiento de conocimiento, el concepto de minería de datos, las operaciones de minería de datos, una descripción general de las técnicas de minería de datos más usadas, los criterios de selección del algoritmo escogido y se finalizará el trabajo con las conclusiones obtenidas.

II. EL DESCUBRIMIENTO DE CONOCIMIENTO

La minería de datos es, en principio, una fase dentro de un proceso global denominado descubrimiento de conocimiento en base de datos (Knowledge Discovery in Database o KDD)[4], aunque generalmente se asocia el concepto de

minería de datos a todo el proceso, en lugar de la fase de extracción de conocimiento. El proceso de KDD es útil en la deserción estudiantil universitaria para obtener conocimientos de las causales generales de ausencia, estudio de variables influyentes según el contexto evaluado, entre otros [15].

KDD se constituye de varias etapas que se ejecutan iterativa e interactivamente. El proceso es no trivial porque incluye acciones de cierta complejidad que involucran la búsqueda de estructuras, modelos y parámetros en la base de datos. Los patrones que se obtienen deben ser válidos con algún grado de certeza, novedosos preferiblemente para el usuario, al que deberán reportar algún tipo de utilidad.

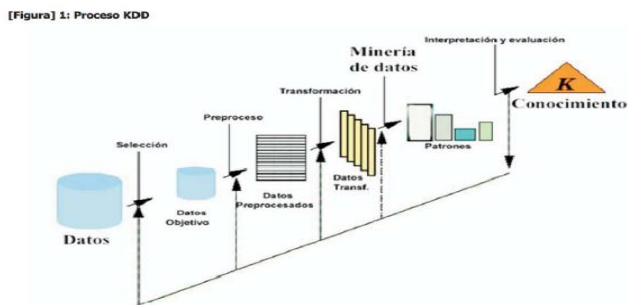


Fig. 1: proceso KDD

El proceso KDD (ver figura 1) comienza con la definición y comprensión de un determinado problema y termina con el análisis de los resultados. Una de las propuestas más ampliamente extendida sobre las etapas o fases componentes del proceso KDD[5] incluye la comprensión del problema, la selección de datos, su limpieza y pre procesamiento, la transformación y aplicación del método de descubrimiento (minería de datos) a utilizar y la interpretación de los patrones obtenidos o análisis de resultados.

A. Minería de datos

La minería de datos es una fase dentro del KDD y se define como “el proceso de extracción de información previamente desconocida, válida y útil de grandes bases de datos y el uso de la información para tomar decisiones cruciales de negocios” [6].

La minería de datos emplea una serie de técnicas las cuales son aplicadas para la solución de diversos problemas [7], sus herramientas predicen futuras tendencias y comportamientos, permitiendo tomar decisiones conducidas por un conocimiento obtenido de los datos. Para conseguir esto hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales.

B. Operaciones de minería de datos

Las cuatro mayores operaciones que se implementan en la minería de datos, son análisis de caminos, segmentación de la base de datos, modelos predictivos y detección de desviación. Hay relación entre las operaciones y los algoritmos, y dependiendo del problema que se quiera resolver y las entradas de datos, se pueden mezclar las operaciones ya que no son mutuamente excluyentes. A continuación se hará una descripción general de tareas.

a) Modelo Predictivo

Es aquel que está relacionado con la experiencia de aprendizaje humano, en la cual se usan observaciones para formar un modelo de las características esenciales de algún fenómeno. La habilidad predictiva es crítica porque ayuda a hacer generalizaciones sobre el mundo que nos rodea y agregar nueva información dentro de un marco general [2].

En el comercio electrónico, por ejemplo, se utiliza este modelo en estrategias de marketing, aprobación de créditos y gerencia de retención de clientes.

b) Segmentación de la Base De Datos

La meta de la segmentación es la partición de la base de datos en segmentos de registros similares. Para particionar los registros se debe tener en cuenta un número de propiedades o características consideradas como homogéneas, los segmentos deben tener homogeneidad y heterogeneidad. La homogeneidad se refiere a que los registros en un segmento están próximos entre sí, donde la proximidad es expresada como una medida dependiente de la distancia de los registros al centro del segmento. La heterogeneidad hace referencia a que los registros en diferentes segmentos no son similares unos de otros de acuerdo con una medida de similaridad, en el comercio, lo anterior se utiliza en la ejecución de estrategias de mercado futuras, tanto on-line como off-line, tales como envío de correo automático a aquellos clientes que se encuentren dentro de un cierto grupo, o presentación de contenidos específicos de publicidad e información según el tipo de cliente.

c) Análisis de Caminos

En contraste con el modelo predictivo y la segmentación de las bases de datos, cuya meta es caracterizar el contenido de la base de datos como un todo, las operaciones de análisis de caminos hacen una búsqueda en la base de datos para establecer relaciones entre registros individuales. Esas relaciones son llamadas asociaciones. Estas asociaciones se

usan para cross selling, target marketing y movimiento de precio común. Por ejemplo, se pueden utilizar para identificar asociaciones relacionadas con las compras a través del tiempo las cuales revelan información sobre la secuencia en la cual los clientes compran productos o servicios, todo esto, para entender perfiles de compra de los clientes a largo plazo y ofrecer promociones más oportunas.

d) *Detección de Desviación*

Las operaciones de desviación son unas de las mejores herramientas para encontrar patrones ocultos en los datos. Dentro de la detección de desviación, la visualización es particularmente útil para percatarse de fenómenos que no se identifican en una muestra relativamente pequeña de los datos. Cuando se usa visualización es importante no tener una hipótesis preconcebida sobre el fenómeno que se está buscando para evitar imprecisiones en los resultados.

Esta operación tiene gran utilidad en las telecomunicaciones, en las cuales se usa un modelo para escanear las millones de transacciones que se realizan diariamente como un esfuerzo para detectar fraudes potenciales.

III. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de minería de datos implementan algoritmos específicos que son usados para realizar las operaciones, a continuación se hará descripción de las técnicas investigadas.

A. *Métodos estadísticos*

El análisis de los datos estadísticos es un sistema de metodologías establecido para la explotación minera de datos. Históricamente, los primeros usos computarizados del análisis de datos fueron desarrollados con la ayuda de la estadística, extendiéndose del análisis de datos unidimensional al análisis de datos multidimensional. La estadística ofrece variedad de métodos para la explotación de minería de datos, incluyendo diversos tipos de regresión y de análisis discriminantes.

B. *Análisis de cluster*

El análisis de clúster es un sistema de metodologías para clasificación automática de muestras en un número de grupos o clases usando una medida de asociación, de modo que las muestras en un grupo sean similares y las muestras que pertenecen a diversos grupos no sean similares. La entrada, para un sistema de análisis de clúster es una muestra y una medida de semejanza (o desemejanza) entre dos muestras.

La salida del análisis de clúster es un número de grupos (clústeres o racimos) en forma de una partición, o una estructura de particiones del grupo de muestra.

C. *Arboles de decisión y reglas de decisión*

Los árboles de decisión y las reglas de decisión son metodologías muy utilizadas en aplicaciones del mundo real de minería de datos como solución de gran alcance a los problemas de clasificación. En general, la clasificación es un proceso de aprendizaje de una función que ingresa un grupo de datos en una de varias clases predefinidas. Cada clasificación está basada en algoritmos de aprendizaje inductivos y se da como entrada a un sistema de muestras que consisten en vectores con valores cualitativos (también llamados vectores de característica) y de una clase correspondiente.

El atractivo de los árboles de decisión se debe a que representan reglas las cuales pueden ser expresadas fácilmente de modo que los seres humanos puedan entenderlas; la capacidad de explicar la razón de una decisión, es crucial.

D. *Reglas de asociación*

Las reglas de asociación son una de las técnicas principales de la explotación de minería de datos y es quizás la más común en el descubrimiento de patrones locales y sistemas de aprendizaje no supervisados [9]. Las técnicas de reglas de asociación recuperan los patrones interesantes posibles en la base de datos, estas se aplican a los conjuntos de transacciones para descubrir todas las asociaciones y correlaciones entre las compras por parte de los clientes.

E. *Detección de fraudes*

Las técnicas de detección de fraude consisten en recopilar datos históricos para construir un modelo de conducta fraudulenta o potencialmente fraudulenta y encontrar instancias similares de esta conducta, estas técnicas se utilizan con mayor frecuencia las aplicaciones de minería de datos de organizaciones como las de cuidado médico, compañías de tarjetas de crédito, prestación de servicios y telecomunicaciones.

IV. CRITERIOS DE SELECCIÓN

Para la selección de una técnica de minería de datos se ha de tener en cuenta una serie de consideraciones previas que afectan al desempeño de la técnica. El entender estas características y su impacto es útil para escoger la técnica que mejor se adecúe a una determinada aplicación. Teniendo en cuenta el entorno se seleccionaron los siguientes criterios [3]:

A. *Criterios*

- 1) Habilidad para manejar datos con ruidos: Las bases de datos a menudo contienen ruido en forma de imprecisiones o inconsistencias. Algunos procesos de validación de datos están mal diseñados y permiten introducir datos incorrectos a los usuarios.
- 2) Habilidad para manejar datos perdidos: es importante darle un manejo apropiado a los datos ya que se pueden producir pérdidas si los datos se obtienen de diferentes fuentes.
- 3) Procesamiento de grandes volúmenes de datos: es importante que la técnica posea la habilidad para manejar gran cantidad de información lo cual permite mayor precisión en el análisis.
- 4) Escalabilidad: esta es una propiedad muy deseable en una técnica de minería de datos para futuras actualizaciones.
- 5) Procesamiento de diferentes tipos de datos: es importante que la técnica seleccionada tenga la capacidad para poder manejar diferentes tipos de datos numéricos, cadenas, etc.
- 6) Capacidad predictiva: Esta característica tiene gran influencia en la efectividad de la técnica de minería porque determina que tan Dueña es para la solución de un problema.
- 7) Facilidad de Operación: La facilidad de integración y operación es otra característica importante para su utilización.
- 8) Capacidad explicativa: Dependiendo de la técnica utilizada, el grado de procesamiento aplicable al dato varía, por tanto, una técnica que sea fácil de entender y que requiera de poco pre procesamiento es más interesante para un usuario final.
- 9) Complejidad de implementación: Es importante que la técnica seleccionada no presente un alto grado de complejidad para su implementación, lo que resulta conveniente para el desarrollo de una herramienta de minería.

B. Algoritmos

a) Algoritmo A priori

El algoritmo a priori trabaja de la siguiente manera: primero busca todos los conjuntos frecuentes unitarios (contando sus ocurrencias directamente en la base de datos), éstos se mezclan para formar los conjuntos de candidatos de 2-itemsets y seleccionan entre ellos los frecuentes. Considerando la propiedad de los conjuntos de ítems frecuentes, vuelve a mezclar éstos últimos y selecciona los frecuentes. Así sucesivamente se repite el proceso hasta que en una iteración no se obtengan conjuntos frecuentes [12].

Este algoritmo asume un orden entre los ítems y utiliza las siguientes notaciones: C_k , para el conjunto de candidatos de k -Itemsets, F_k , para el conjunto frecuente de k -itemsets y

asociado a cada itemset se encuentra el campo count para almacenar el soporte de dicho itemset.

b) Algoritmo DHP (Poda y hashing directa)

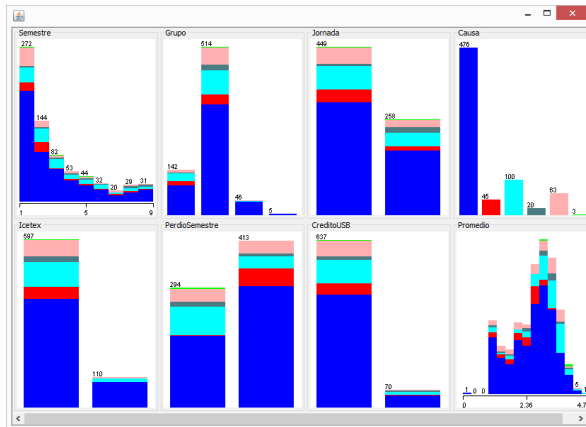
En el algoritmo de poda y hashing directa (DHP, Direct hashing and Pruning) se emplea una técnica de hash para eliminar todos los conjuntos de ítems innecesarios para la generación del próximo conjunto de ítems candidato. Cada $(k+1)$ -Itemset es añadido a una tabla hash en un valor hash dependiente de las ocurrencias en la base de datos de los conjuntos candidatos de k elementos que lo formaron, o sea, dependiente del soporte de los conjuntos candidatos de k elementos. Estas ocurrencias son contadas explorando en las transacciones de la base de datos. Si el soporte asociado a un valor hash es menor que el soporte mínimo, entonces, todos los conjuntos de ítems de $k+1$ elementos con este valor hash no serán incluidos entre los candidatos de $k+1$ elementos en la próxima iteración.

c) Algoritmo Partition

Este algoritmo en primer lugar propone particionar la base de datos en tantas partes como fueran necesarias para que todas las transacciones en cada partición sean alojadas en la memoria principal operativa. En contraste con otros algoritmos, este recorre la base de datos sólo dos veces. La primera vez, cada partición es minada independientemente para encontrar todos los conjuntos de ítems frecuentes en la partición y luego se mezclan éstos para generar el conjunto de los conjuntos de ítems candidatos. En la segunda iteración se cuenta la ocurrencia de cada candidato, aquellos cuyo soporte es mayor que el mínimo soporte especificado, se retienen como conjuntos frecuentes. Este algoritmo emplea el mecanismo de Intersección entre conjuntos para determinar el soporte de dichos conjuntos, en este caso, cada ítem en una partición mantiene la lista de los identificadores de las transacciones que contienen a dicho ítem.

d) Algoritmo Eclat

EL algoritmo Eclat reduce la cantidad de operaciones de E/S, aunque atraviesa la base de datos sólo una vez. Se basa en realizar un agrupamiento (clustering) entre los ítems para aproximarse al conjunto de ítems frecuentes maximales y luego emplea algoritmos eficientes para generar los ítems frecuentes contenidos en cada grupo. Para el agrupamiento hay dos métodos que son empleados después de descubrir los conjuntos frecuentes de dos elementos:



1. Por clases de equivalencia: esta técnica agrupa los itemsets que tienen el primer ítem igual.
2. Por la búsqueda de cliques maximales: se genera un grafo de equivalencia cuyos nodos son los ítems y los arcos conectan los ítems de los 2-itemsets frecuentes. Se agrupan los ítems por aquellos que forman cliques maximales.

Con el primero de los métodos se generan todos los conjuntos de ítems frecuentes, con el segundo, se generan sólo los conjuntos frecuentes maximales, los restantes conjuntos frecuentes son subconjuntos de éstos

e) Algoritmo FP-Growth

Este algoritmo está basado en una representación de árbol de prefijos de una base de datos de transacciones dada (llamada FP-Tree), el cual puede almacenar considerables cantidades de transacciones en la memoria. La idea básica del algoritmo FP-Growth puede ser descrita como un esquema de eliminación recursiva: en un primer paso de preprocesamiento se borran todos los ítems de las transacciones que no son frecuentes individualmente o no aparecen en el mínimo soporte de transacciones, luego se seleccionan todas las transacciones que contienen al menos un ítem frecuente (al menos una cantidad de esos que son frecuentes). Luego se hace el proceso recursivo para obtener una base de datos reducida. Al retorno, se remueven los ítems procesados de la base de datos de transacciones en la memoria y se empieza otra vez, y así con el siguiente ítem frecuente [13].

V. CRITERIOS DE SELECCIÓN DEL ALGORITMO

Para la selección del algoritmo de Reglas de Asociación hay que tener en cuenta una serie de consideraciones que tienen impacto en la solución que se desea brindar

• **Inserciones a la base de datos:** Es importante que los algoritmos minimicen el recorrido por la base de datos, pues el número de reglas crece exponencialmente con el número de

ítems considerados, lo cual afecta el rendimiento del algoritmo cuando se accesa constantemente a la base de datos.

- **Costo computacional:** es importante que el algoritmo no realice un gran número de operaciones que agoten los recursos de máquina.
- **Tiempo de Ejecución:** se desea que el tiempo utilizado para la generación de reglas sea razonable.
- **Rendimiento:** es importante que el algoritmo realice las operaciones y procesos de forma eficiente.

La investigación de técnicas de Minería de datos aplicadas a problemas específicos permitió delimitar los requerimientos del proyecto y el uso de diferentes algoritmos rápidos y eficientes permitiendo la realización del proceso de minería de datos de forma rápida.

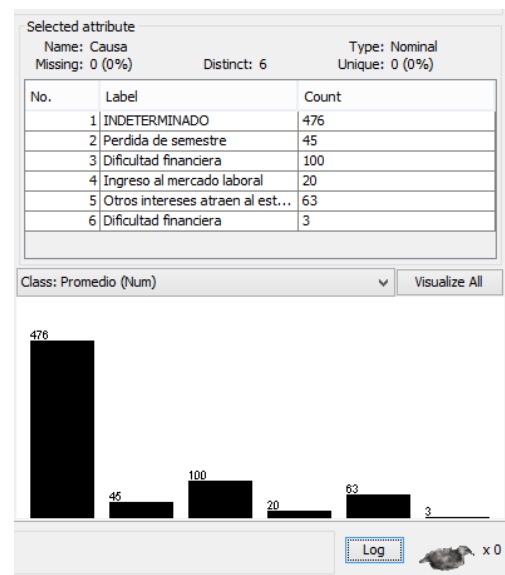
VI. APLICACIÓN DE LA MINERÍA DE DATOS Y ANÁLISIS DE DESERCIÓN ESTUDIANTES USB

Para el estudio de esta problemática de deserción estudiantil en la Universidad Simón Bolívar, se agrupó las causales en 5 variables definidas que son:

- Perdida de semestre
- Dificultad financiera
- Ingreso al mercado laboral
- Otros intereses atraen al estudiante
- Indeterminado.

Para ello hemos extraído una población considerable, abarcando desde el primer al décimo semestre del programa de Ingeniería de Sistemas, entre los períodos académicos 2007-2012.

Podemos destacar que la principal causa de deserción de los estudiantes es Indeterminada, lo cual debe generar un plan de acción, para realizar mejores filtros en la búsqueda de estas causales, a través de diferentes estrategias de recopilación de



la información necesaria. Estos registros son de suma importancia para las decisiones administrativas y financieras.

Seguidamente encontramos la dificultad financiera, otros intereses atraen al estudiante, pérdida de semestre e ingreso al mercado laboral.

A continuación una vista general de las variables que se analizan, conjuntamente:

Weka nos permite clasificar los datos de acuerdo a los algoritmos predeterminados de análisis. Aquí se aprecian las características de la población de estudio (muestra): semestre, grupo, jornada, causa, préstamo con Icetex, perdió el semestre, crédito con la Universidad Simón Bolívar y promedio académico.

Weka nos permite agrupar un rubro de conclusiones pertinentes junto con el desarrollo de las metodologías de minería de datos anteriormente explicadas.

Una conclusión certera abre camino a estrategias de grupo para disminuir los niveles de deserción, por ejemplo, de la herramienta podemos definir como dato final de un análisis que la causa indeterminada de deserción se da en estudiantes de primer semestre académico, en el grupo A, jornada diurna, la mayoría de los estudiantes no poseían crédito con Icetex ni con la USB, con pérdidas de semestres, y cuyo promedio general oscila entre los 2.40 y 3,0.

VII. CONCLUSIONES

- La investigación de las diferentes técnicas de minería de datos y su empleo en la solución de diversos tipos de problemas de análisis de información nos ayudan a tener un conocimiento general del tema para desarrollar trabajos futuros en otras áreas de conocimiento.

- Para la escogencia de una técnica de minería de datos es necesario entender las necesidades propias del trabajo a desarrollar y tener en cuenta consideraciones y criterios para seleccionar una técnica adecuada que satisfaga los requerimientos propios del mismo.

- Lo importante del desarrollo de la investigación es lograr que la Universidad Simón Bolívar genere estrategias administrativas para mitigar de alguna manera las variables que afectan directamente en la deserción estudiantil, logrando sensibilizar a los directivos y otorgarles información base, para la toma de decisiones que competan al curso de la población estudiantil en cuestión.

- Planes de acción: Bono de matrícula, Monitoria social, Vinculación laboral, Reliquidación de matrícula, Estrategia comunicativa, Educación de las familias y acudiente, Intervención familiar, Estrategias de acompañamiento académico, Acompañamiento individual por psicología.

VIII. REFERENCIAS

- [1] Jose Alberto Castañerda Garcia Miguel Angel Rodriguez Molina "La minería de datos como herramienta de Marketing: Delimitación y Evaluación del resultado" Facultad de CC. EE
- [2] Departamento de Comercialización e Investigación de mercados CABENA, Hadjinjan., STADLER, Verhees Zanasi., Discovering Data mining From Concept To Implementation. Prentice Hall PTR. Upper Saddle River. 1998 - 224 paginas.
- [3] CABENA, Hadjinjan., STADLER, Verhees Zanasi., Discovering Data mining From Concept To Implementation. Prentice Hall PTR. Upper Saddle River. 1998 - 224 paginas.
- [4] Alejandro Amat Bedmar - "Ingeniería De Conocimiento Minería De Datos Empresariales" - M.S. / E.T.S. Ingeniería Informática de la Universidad de Granada - 2005.
- [5] Garlan Dv «Software Architecture: a Roadmap,» The Future of Software Engineering, ACM Press, pp.91-101,2000. J. Zhao, «Applying Slicing Techmque to Software Architectures," Proc. Fourth IEEE International Conference on Engineering o(Complex Computer Systems (ICECCS98), pp.87-98, Agosto - 1998.
- [6] Fayyad U, Piatetsky-Shapiro G, Smyth P.,Discovery and Data Mining. AAAI Press/The MIT Press -1996
- [7] KENNETH C. LAUDON, Jane P. Administración de la Información y toma de decisiones, Resúmenes de los principales capítulos del libro, Management Information Systems Organization and Technology. Documento. Universidad de Taparaca. Chile - 2005.
- [8] KIMBALL, Ralph., ROSS, Margy. The Data Warehouse Toolkit The Complete Guide to Dimensional Modeling. Libro. McGraw Hill. Second Edition 2002 - 464 paginas.
- [9] Piatetsky-Shapiro G. Report on the AAAI-91 Workshop on Knowledge Discovery in Databases, IEEE Expert 6(5), pp.74-76 Octubre -1991.
- [10] Mehmed Kantardzic - Data Mining: Concepts, Models, Methods, and Algorithms - John Wiley & Sons -2003 -343 paginas.
- [11] JaweJ Han, Micheline Kamber - Data Mining: Concepts and Techniques Simón Fraser University Morgan Kaufmann Publishers-2002 - 550 paginas.
- [12] Dittl T. "DikLuvainy Knowludyw in DdLd. An Introduction to Data Mining" - John Wiley & Sons - 2004 -240 paginas.
- [13] Roxana Darvger Mercaderes - Rafael Berlanga Llavori- Informe técnico "Búsqueda de Reglas de Asociación en bases de datos y colecciones de textos" - Departamento de Computación, Universidad de Oriente, Santiago de Cuba 2001.
- [14] Christian Borgelt-Department of Knowledge Processing and Language Engineering -School of Computer Science, Otto von Guericke - University of Magdeburg 2005.
- [15] J.RAYPORT JEFFREY, R, JAWORSKI BERNARD J. Comercio Electrónico. McGraw Hill. - 2000 - 456 paginas.
- [16] NUÑEZ, Fernando Alberto., LUGONES, Fernando Alberto., Modelos de Negocios en Internet visión poscrisis. McGraw/Hill. 2001- 384 paginas.
- [17] Waikato ML Group. The Waikato environment for knowledge analysis. <http://www.cs.waikato.ac.nz/ml/weka> - 2007
- [18] Faculty of Computer and Slovenia Information Science, University of Ljubliana. Orange, fruitful and fun. <http://www.ailab.si/orange> - 2007.
- [19] R Rakotomalala. Tanagra project. <http://chiroubte.univ-lyon2.fr/ricco/tanagra/en/tanagra.html> • 2007.
- [20] The CRISP-DM consortium. CRISP-DM Step by step data mining guide. -2000 Documento. <http://www.crisp-dm.org/CRISPWP-0800.pdf> ,2007.