Herramientas y funcionalidades de minería de datos en Oracle

Por Ing. Alexei Rodríguez Méndez, Técnico en Sistema de Computación,
Gerencia de Innovación y Desarrollo, UNTISW, ETECSA
alexei.rodriguez@etecsa.cu

Introducción

Hoy día la informatización de la sociedad es toda una realidad, millones de sistemas informáticos han sido desarrollados y, de una forma u otra, apoyan, ejecutan o controlan las actividades y procesos diarios. El empleo estratégico de la información ha adquirido nuevos matices, donde quien la posee, tiene una gran ventaja y posibilidades de éxito.

El Descubrimiento del Conocimiento en Bases de Datos —del inglés, *Knowledge Discovery in Databases* (KDD)— consiste en el proceso de extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos. El KDD posee varias etapas donde la más importante es la Minería de Datos que se basa en la aplicación de técnicas de diferentes campos como la inteligencia artificial y la estadística a grandes volúmenes de datos con el objetivo de encontrar patrones y relaciones no conocidas y, a veces, insospechadas [7]. A través de ella, pueden explicarse comportamientos que serían muy difíciles de diagnosticar con otras técnicas tradicionales. La realización de proyectos de minería de datos tiene sus peculiaridades, existen metodologías, herramientas y numerosos estudios los cuales apoyan a los especialistas de esta rama en la ejecución de tales tareas.

En este artículo se presentan los principales conceptos relacionados con la Minería de Datos, con énfasis particular en las herramientas y funcionalidades que aporta Oracle en tal sentido.

Minería de Datos

Para hablar de Minería de Datos primero es necesario mencionar el KDD, este se define como un proceso de extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos [4]. El KDD cuenta con varias etapas —compresión del negocio, selección de datos, limpieza y preprocesamiento, transformación, minería de datos, evaluación e interpretación de los resultados—.

La Minería de Datos es una etapa del KDD [4], e incluye el análisis de grandes volúmenes de datos, con el objetivo de encontrar relaciones no conocidas y resumirlas de forma novedosa y útil para los dueños de la información. Los resultados son conocidos como patrones o modelos [5]. Los procesos pueden ser automáticos o semiautomáticos y los patrones descubiertos deben ser significativos y ventajosos para el interesado [10]. Las técnicas de minería de datos permiten obtener predicciones válidas [1].

Existe un dilema dado fundamentalmente por los términos de KDD y Minería de Datos (MD). Algunos autores consideran que el primero

es más amplio y abarcador que el segundo, donde MD sólo se refiere al conjunto de algoritmos y métodos empleados para extraer el conocimiento y forma parte del proceso del KDD. Por otra parte, es común encontrar el concepto de minería de datos como un símil del KDD.

El término Minería de Datos es más empleado por los estadísticos, analistas de datos y los sistemas de administración de la información. KDD, por su parte, tiene seguidores en los estudiosos de campos de la inteligencia artificial y las máquinas de aprendizaje.

Metodologías de Minería de Datos

Enfrentar un proyecto de Minería de Datos requiere de experiencia, capacitación pero, sobre todo, planificación y organización. Las metodologías para el desarrollo de software como RUP [6], XP [11], han logrado estandarizar los procesos de software. La Minería de Datos no puede tratarse de forma tradicional, tiene sus propias características, de ahí que existen metodologías para ejecutar estos proyectos. Entre ellas las más empleadas son CRISP-DM —del inglés, *CRoss-Industry Standard Process for Data Mining*— y SEMMA —del inglés, *Sample, Explore, Modify, Model, Assess*—.

Metodología CRISP-DM

La metodología CRISP-DM fue creada en 1996 cuando un importante consorcio de empresas europeas —NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda)— unieron sus recursos para el desarrollo de esta metodología de libre distribución. CRISP-DM ha tenido éxito porque está basada en la práctica, en experiencias reales de cómo los expertos realizan los proyectos de Minería de Datos [9].

Es una metodología con propósitos generales para cualquier proyecto de MD. Plantea ideas que deben parametrizarse para cada entorno de ejecución, desechando algunas cosas y adicionando otras, según sea la naturaleza y los objetivos del proyecto. Propone modelos genéricos que deben ser adaptados: a esta acción se le denomina mapear el modelo.

CRISP-DM plantea cuatro niveles de abstracción durante un proyecto de DM, organizados de forma jerárquica en tareas que van desde las generales hasta las específicas —fases, tareas genéricas, tareas específicas, instancias de procesos—. También propone un modelo de referencia compuesto por 6 fases relacionadas entre sí y que interactúan de forma cíclica como muestran la figura 1.

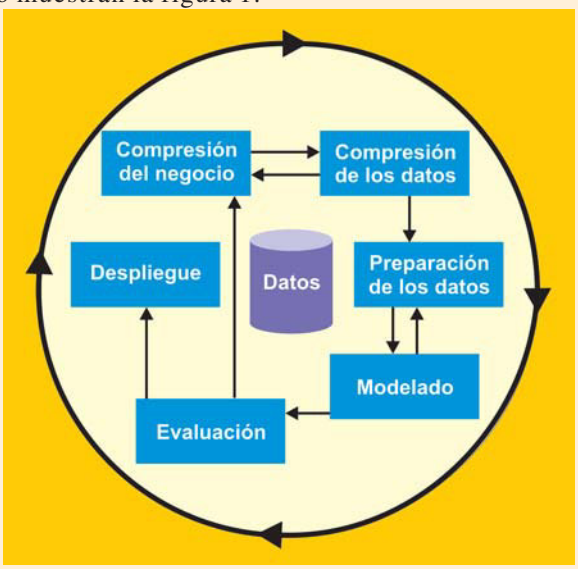


Figura 1 Fases del modelo de referencia CRISP-DM

Metodología SEMMA

Esta metodología es desarrollada por SAS, empresa a la vanguardia en temas de minería de datos e inteligencia de negocio —*business intelligence*—. SEMMA define una organización lógica de actividades que se ejecutan en el empleo de SAS Enterprise Miner para realizar proyectos de minería de datos. Su nombre está formado por las iniciales de las etapas que propone: ***S**ample* (Muestreo), ***E**xplore* (Exploración), ***M**odif.* (Modificación), ***M**odel* (Modelado), ***A**sses* (Evaluación) [2] (Figura 2).

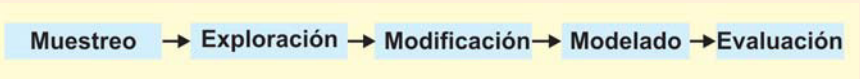


Figura 2 Etapas de SEMMA

Comparación entre CRISP-DM y SEMMA

Ambas metodologías son efectivas en un proyecto de minería, estructuran las tareas en fases donde básicamente se manifiestan las siguientes etapas: extracción de los datos —preparación—modelado—evaluación—despliegue. Estas etapas están interrelacionadas, por ejemplo, al obtener un modelo es posible que haya que realizar una nueva preparación de los datos u otra selección de los mismos.

SEMMA es más limitada en cuanto a la comprensión del problema desde el punto de vista empresarial. Comienza con la realización de una extracción de los datos, mientras que CRISP-DM propone una lógica más abarcadora, es decir, entiende el negocio y por qué es necesario y dónde realizar un proyecto de DM.

CRISP-DM es publicada y distribuida libremente, puede ser implementada por cualquier interesado en el tema. SEMMA sólo muestra sus aspectos generales y los acopla a su producto de minería Enterprise Miner.

SAS actualmente ha planteado que SEMMA no es una metodología, sino una organización de pasos para hacer minería con su producto de DM.

Minería de Datos en Oracle

Oracle Corporation en una compañía líder en el mundo en materia de base de datos y otras aplicaciones. Su producto insignia es el motor de Base de Datos (BD) Oracle. En su edición Enterprise, a partir de la versión 9i, incluye técnicas de minería de datos concebidas en *Oracle Data Mining* (ODM). Estas funcionalidades están completamente embebidas en el propio motor de la base de datos, por lo que no requieren de procesos de instalación extra.

ODM tiene sus orígenes en el producto Darwin desarrollado por Thinkign Machines Corp., que fue adquirido por Oracle en 1999 [3]. ODM es una infraestructura que permite construir aplicaciones robustas de minería sin tener que recurrir a softwares de terceros. Los procesos de extracción del conocimiento se simplifican, debido a que se elimina la necesidad de movimientos de los datos para su análisis. Todas las actividades de preparación, creación de modelos y análisis se realizan en la BD, influyendo directamente en un aumento de la productividad y efectividad de los resultados.

La integración de ODM con la base de datos se logra a través de las interfaces Java y PL/SQL. La interfaz Java facilita la creación de aplicaciones tipo java que pretendan hacer minería en Oracle. Por otra parte, existen las interfaces PL/SQL DBMS_DATA_MINING y DBMS_DATA_MINING_TRANSFORM, para el acceso a las técnicas de minería en aplicaciones PL/SQL. Es importante señalar que, aunque ambas interfaces en un principio permiten las mismas funciones, existen diferencias entre ellas.

Las funciones de Minería de Datos están basadas en dos tipos de aprendizajes —supervisados y no supervisados—. Los supervisados son usados generalmente para predecir valores y son implementados en modelos predictivos. Por otra parte, los no supervisados se emplean en problemas donde no han sido definidos objetivos o variables a determinar. Estos últimos pueden arrojar resultados insospechados.

Entre las etapas propuestas de un proyecto de minería de datos se encuentra la obtención del modelo, para ello se utilizan varios algoritmos provenientes, principalmente, de la estadística e inteligencia artificial. ODM implementa muchos de estos algoritmos conocidos y aporta sus variantes. Las funciones de minería que soporta ODM son las siguientes [8]:

- Modelos predictivos —aprendizaje supervisado—: clasificación, regresión, importancia de atributo.
- Modelos descriptivos —aprendizaje no supervisado—: segmentación, modelos de asociación, extracción de patrones.

ODM también tiene soporte para aplicaciones de minería de texto, así como funciones específicas para el campo de la bioinformática a través de la herramienta BLAST. También, implementa algoritmos diversos, algunos tradicionales y otros propietarios, a continuación en la tabla 1 se exponen los tipos de problemas y los algoritmos que pueden ser usados.

Algoritmos Problemas	Árboles Decisión	*Adaptative Bayes Network*	*Naive Bayes*	SVM	*One Class SVM*	*A priori*	Matriz factorización no negativa	*Min Desc Length*	*K-Means*	*O-Cluster*
Detección Anormalidades					X					
Reglas Asociación						X				
Importancia Atributos								X		
Clasificación	X	X	X	X						
Clustering									X	X
Extracción Patrones							X			
Regresión				X						

Tabla 1 Tipos de problemas y algoritmos

Oracle Data Miner

Oracle Data Miner es una herramienta de minería de datos desarrollada por Oracle. Es multiplataforma y utiliza las interfaces de ODM. Cuenta con una interfaz de usuario sencilla e intuitiva. Implementa variadas funciones de preprocesamiento de los datos de manera visual completamente, el usuario no necesita conocer lenguaje alguno de consultas para realizar estas actividades. Es parametrizable y permite exportar los resultados para ficheros csv, así como la importación de fuentes de datos externas a través de ficheros textos.

Para la visualización de los resultados Oracle Data Miner propone procedimientos estándares para cada tipo de actividad de minería, donde puede comprobarse gráficamente la calidad de los modelos obtenidos, así como otros indicativos de evaluación como la matriz de costo.

Una característica de reciente incorporación en la versión 10.2.2 es la exportación de los paquetes PL/SQL hacia la BD Oracle para ejecutar las actividades de minería realizadas con Oracle Data Miner, pero desde otras aplicaciones que empleen la BD. A través de esta importante característica, se facilita, en gran medida, la implementación de aplicaciones de minería de datos utilizando la BD Oracle.

Empleo de Oracle Data Miner

Oracle Data Miner soporta gran cantidad de tareas de minería de datos, desde sus fases iniciales de extracción, transformación hasta la evaluación de los resultados y el despliegue.

Funcionalidades para la preparación y transformación de los datos

La exploración de los datos puede ser realizada a través de resúmenes para registros simples o multirregistros, donde en cada caso se muestra, mediante un histograma —como muestra la figura 3—, la distribución de los datos; en caso de que el dominio sea muy amplio, se emplean técnicas como la discretización.

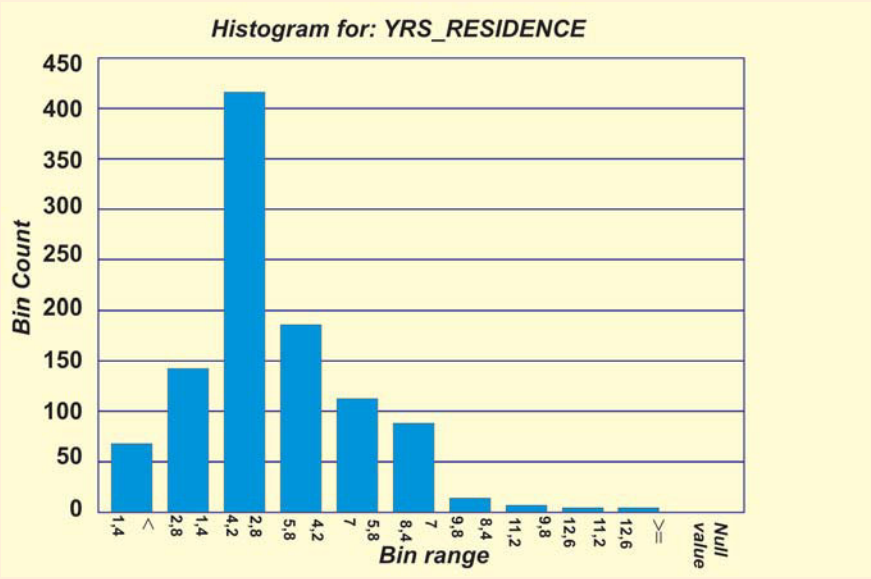


Figura 3 Histograma para el campo años de residencia

Para la transformación de los datos y su exploración, existen varias funciones, entre ellas: agregación, campos computados, discretización, filtrado de campos, valores perdidos, normalización, tratamientos de puntos fuera de rango, recodificación —permite sustituir valores por otros según criterios de entrada—, selección aleatoria, selección aleatoria por capas, división, división por capas (Figura 4).

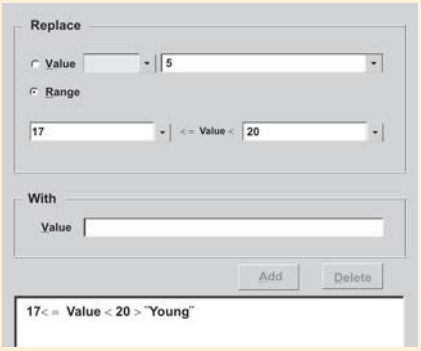


Figura 4 Transformación por recodificación

Con las funciones anteriores pueden ejecutarse variadas actividades de preparación de los datos sin necesidad de ser un conocedor de la sintaxis sql, ni de las particularidades de Oracle. Además, existen funcionalidades relacionadas con la explicación de los datos, donde el especialista puede conocer qué influencia tiene un determinado campo sobre el resultado final, mostrado también en forma de histograma pero horizontalmente.

Construcción del modelo

Para esta etapa del desarrollo de un proyecto de minería de datos, la herramienta posee diversas características. La secuencia de pasos varía en dependencia del tipo de actividad desarrollada y el algoritmo empleado, aunque existen puntos comunes para todos los proyectos que puedan desarrollarse. A continuación se explican las acciones a ejecutar en un problema típico de clasificación.

El primer paso es seleccionar el tipo de tarea que se desea ejecutar —asociación, clasificación, regresión, detección de anomalías, segmentación, importancia de atributos, etc.—, en este caso clasificación y el tipo de algoritmo a emplear. Después se selecciona la fuente de datos, especificando los campos a tener en cuenta y un identificador para cada registro; adicionalmente se permite incluir otras fuentes de datos, lo cual no debe ser necesario si se ejecutó una preparación adecuada. El paso siguiente consiste en señalar cuál es la variable obje-

tivo del proyecto actual y su valor preferido —el algoritmo tratará de maximizar la ocurrencia del valor dado— (Figura 5).

Figura 5 Formulario de selección de variable objetivo y campos a incluir en el modelo

Una vez realizado los pasos anteriores, la actividad está lista para ser procesada y obtener el modelo correspondiente. La herramienta ejecuta una serie de tareas predefinidas que pueden ser personalizadas si se desea. Entre ellas se encuentran: una segunda selección de los datos, por defecto no está activada porque se supone que se hayan ejecutado acciones anteriores con esos objetivos. Discretización del dominio, está activada por defecto aunque puede ser desactivada si se realizó esta acción en la fase de preparación, en caso contrario, no es recomendable debido a que puede que la corrida del modelo se torne muy lenta. Particionamiento de los datos, se especifica qué sección será destinada a la construcción del modelo y cuál a la prueba del mismo, una razón recomendada es (60–40) respectivamente. Parámetros de la construcción del modelo entre los que se encuentran: metas de precisión —precisión máxima promedio, precisión máxima global— y parámetros propios para cada algoritmo.

Una vez corrido el modelo, los resultados de cada etapa o pasos vistos anteriormente pueden ser consultados en la pantalla de resultado de la actividad que aparece en la figura 6.

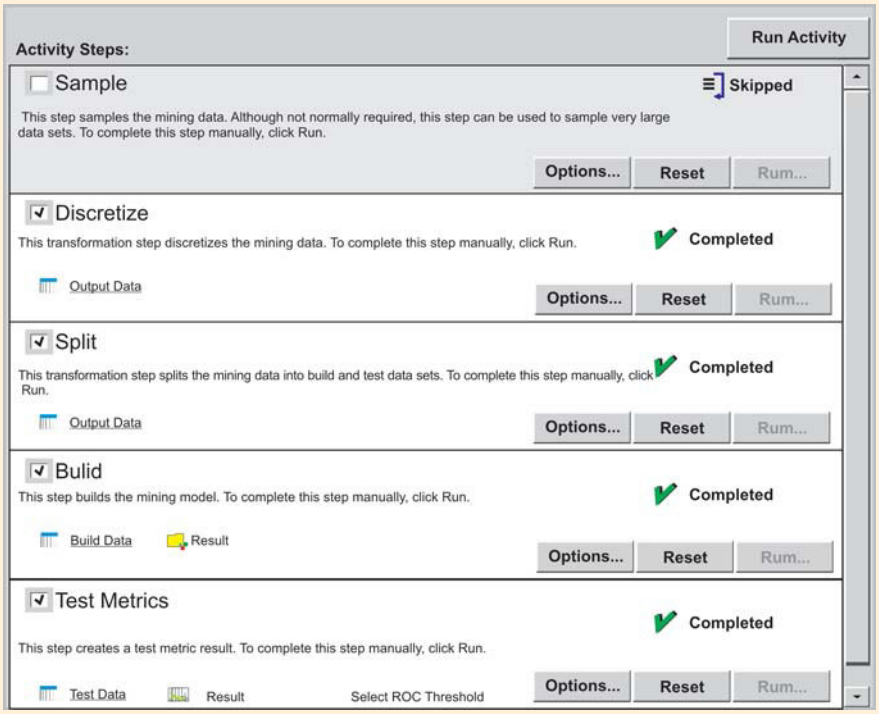
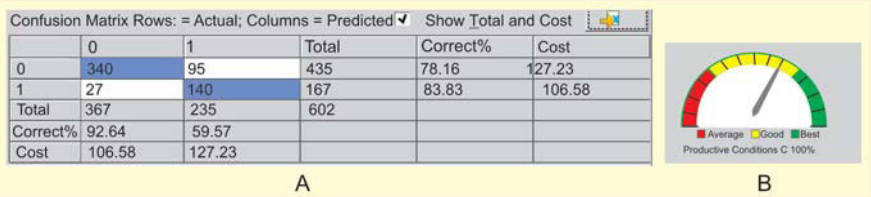


Figura 6 Pantalla de resultados de la corrida del modelo

La tarea de evaluación de los resultados o métricas de resultados —*Test Metrics*— posee los datos más interesantes, porque Oracle Data Miner brinda varios artefactos de visualización de los resultados donde, de una forma muy interesante, el especialista conoce la efectividad del modelo hallado.

La gráfica de confianza de la predicción (Figura 7b), indica la calidad del modelo. La zona verde significa mejores modelos, pero se considera un buen resultado si la aguja está en la zona derecha del reloj.

La página de precisión (Figura 7a) muestra varias interpretaciones de la exactitud del modelo. Partiendo de que en los datos empleados se conoce el valor real de la variable objetivo, se construyen gráficas donde puede verificarse en qué cuantía el modelo predijo los resultados reales. Otro aspecto significativo es la matriz de confusión, a través de ella se conoce el costo de hacer una mala predicción y así tomar decisiones.



	0	1	Total	Correct%	Cost
0	340	95	435	78.16	127.23
1	27	140	167	83.83	106.58
Total	367	235	602		
Correct%	92.64	59.57			
Cost	106.58	127.23			

Figura 7 Matriz de confusión y gráfica de confianza

La página de ROC muestra una gráfica muy útil para determinar posibles cambios en el modelo, responde preguntas como: ¿qué pasa si se cambia *x* parámetro? La gráfica muestra la relación entre los resultados positivos verdaderos y los falsos positivos en los segmentos de datos, por defectos los datos son ordenados por probabilidad y divididos en 10 partes. El usuario puede desplazar la línea roja hasta lograr la relación deseada de positivos y negativos, según sea el interés (Figura 8a).

La página de acumulación de probabilidades (Figura 8b), muestra otra interpretación de los resultados por probabilidades, las observaciones realizadas aquí son del tipo ¿qué cantidad de veces el modelo es mejor que una solución aleatoria del problema visto?

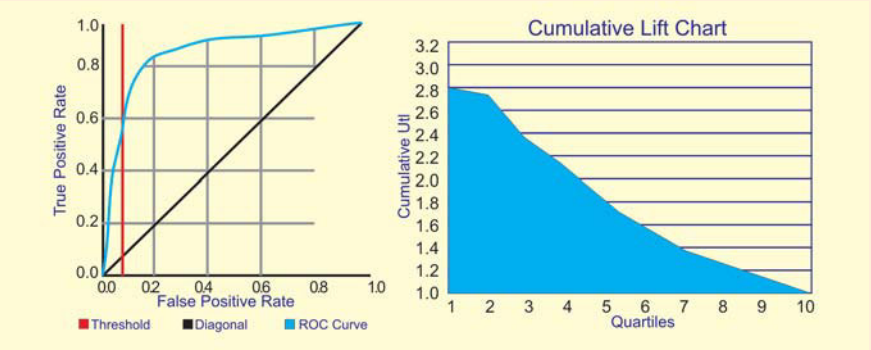


Figura 8 Gráficas de ROC y de acumulación de probabilidades

Oracle Data Miner implementa utilitarios de exportación de los modelos a paquetes PL/SQL en una base de datos Oracle, de esta forma el especialista informático puede emplearlos en aplicaciones más personalizadas. También permite publicar los resultados en un servidor Oracle Discover y al formato PMML.

Conclusiones

Actualmente, la Minería de Datos es un campo aún en desarrollo, nuevas tecnologías se abren camino cada día, Oracle constituye con sus productos de minería una solución ventajosa si los datos yacen en este tipo de base de datos. ODM es muy factible de emplear con resultados rápidos y confiables que permiten escalar en profesionalidad y eficacia.

Referencias bibliográficas

[1] Eldestein, Hebert A. *Introduction to Data Mining and Knowledge Discovery. Third Edition*. USA: Two Cross Corporation, 2005.

[2] E-Intelligence, *S. Finding the Solution to Data Mining. A Map of the Features and Components of SAS ® Enterprise Miner TM Software Version* 4.1, 2000. Disponible en: http://www.sas.com (Consulta: 20/11/2006).

[3] "Thinking Machines Purchased by Oracle", Disponible en: http://query.nytimes.com/gst/ (Consulta: 21/06/2006).

[4] Fayad, U.M., Piatetsky-Shapiro, G. & Smyth, P. "From Data Mining to Knowledge Discovery in Databases". *AAAI 97*, no. (1996): 18. Disponible en http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. (Consulta: 08/05/2006).

[5] Hand, David; Mannila, Heikki; y Smyth, Padhraic. *Principles of Data Mining*. E.U.: MIT Press, 2001, 546 págs.

[6] Jacobson, Ivar; Booch, Grady; y Rumbaugh, James. *El proceso unificado de desarrollo de software*. USA: Addison Wesley, 2006.

[7] Méndez, A. R. *Empleo de técnicas de Minería de Datos con soporte Oracle en apoyo a la toma de decisiones relacionado con fraude en las reclamaciones telefónicas*. La Habana: CUJAE, 2007, 56 págs.

[8] Oracle_Corporation. *Oracle Data Mining Concepts*, 10g Release 1 (10.1), Corporation Oracle, 2003. Part No. B10698-01: 118. Disponible en: http://www.oracle.com/technology/documentation/datamining.html. (Consulta: 15/11/2006).

[9] Chapman, Pete (NCR), J. C. S., Kerber, Randy (NCR), Khabaza, Thomas (SPSS), Reinartz, Thomas (DaimlerchrChrysler), Shearer, Colin (SPSS) and Wirth, Rüdiger (DaimlerchrChrysler). *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, 2000. Disponible en: http://www.crisp-dm.org. (Consulta: 13/10/2006).

[10] Witten Ian H.; Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques. Second Edition*. San Francisco C.A.: Elseiver Inc., 2005, 558 pp.

[11] "XP. What is Extreme Programming?" (2006). Disponible en: http://www.extremeprogramming.org/what.html. (Consulta: 25/06/2007).