INTEGRANDO ÁREAS DISCIPLINARES EN UN DISEÑO CURRICULAR

Mag. Graciela Beguerí, Mag. Alejandra Malberti, Mag. Raúl O. Klenzi
Instituto de Informática – Departamento de Informática
Facultad de Ciencias Exactas, Físicas y Naturales – Universidad Nacional de San Juan
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", San Juan
{grabeda, amalberti, rauloscarklenzi}@gmail.com

Resumen

Este artículo muestra como se pueden vincular áreas disciplinares y a la vez incorporar, integrar y explotar simultáneamente distintas herramientas libres, provenientes de los aportes de las Tecnologías de la Información y la Comunicación –TIC-, en un diseño curricular en general y en particular en las carreras de grado del Departamento de Informática de la Facultad de Ciencias Exactas, Físicas y Naturales – Universidad Nacional de San Juan. FCEFN-UNSJ.
Se toma como punto central el tema Minería de Datos, por ser éste un campo multidisciplinar que se ha nutrido de diferentes áreas conceptuales tales como, estadística, aprendizaje automático y bases de datos entre otras. Esta tarea de vinculación conlleva una necesaria articulación vertical y horizontal de las asignaturas que imparten los contenidos, lo que no sólo debe evidenciarse de manera explícita en los planes de estudio sino también en el proceso de enseñanza aprendizaje.
La presente propuesta surge desde la experiencia alcanzada por docentes pertenecientes a las áreas curriculares de Ciencias Básicas, Algoritmos y Lenguajes, Teoría de la Computación e Ingeniería de Software-Base de Datos-Sistemas de Información, de las carreras de Informática, e integrantes de sucesivos proyectos de investigación en el área temática de minería de datos, desarrollados en el ámbito de la FCEFN.

Palabras clave: Estadística – Minería de Datos – Informática – Proceso de enseñanza aprendizaje

Introducción

El avance de la tecnología de Minería de Datos-MD ha permitido salvar algunos factores que desalentaban su aplicación, como ser: los productos comerciales eran muy costosos y en general los análisis estadísticos y de MD eran abordados mediante el uso inconexo de diferentes productos.
Paralelamente, los avances en las comunicaciones, la evolución del hardware, y de los entornos open source han influido en el aumento de repositorios (sitios centralizados que pueden albergar todo tipo de materiales digitales, desde ficheros textuales a audiovisuales, con el propósito no sólo de facilitar su alojamiento sino de garantizar su preservación, acceso y distribución), los que conjuntamente con los sistemas de bases de datos hacen que existan grandes cantidades de datos en diferentes formatos, susceptibles de ser tratados con estrategias de MD.

Hoy en día existe la posibilidad de contar con una enorme diversidad de software de código libre y abierto que caracteriza al trabajo en repositorios de datos y que puede ser utilizado tanto académicamente como comercialmente, abriendo así oportunidades educativas y laborales [20] [21].
Desde lo conceptual se puede observar que numerosa bibliografía de MD contiene explícitamente capítulos de Estadística, presentándose la misma situación de forma inversa [3] [6] [7] [8] [9] [10] [11] [12] [14] [17] [19].
Si bien los algoritmos con que se vale MD se nutren de modelos estadísticos, aún persiste una desvinculación académica y práctica entre ambos tópicos.

Diversos trabajos expresan que:

La minería de datos (en inglés, data mining) se define como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos. En la actual sociedad de la información, donde día a día se multiplica la cantidad de datos almacenados casi de forma exponencial, la minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización. La minería de datos se define también como el análisis y descubrimiento de conocimiento a partir de datos.

La minería de datos hace uso de todas las técnicas que puedan aportar información útil, desde un sencillo análisis gráfico, pasando por métodos estadísticos más o menos complejos, complementados con métodos y algoritmos del campo de la inteligencia artificial y el aprendizaje automático que resuelven problemas típicos de agrupamiento automático, clasificación, predicción de valores, detección de patrones, asociación de atributos, etc. Es, por tanto, un campo multidisciplinar que cubre numerosas áreas y se aborda desde múltiples puntos de vista, como la estadística, la informática (cálculo automático) o la ingeniería [7] [17].

El proceso de acreditación al que se ven sometidas las diferentes carreras de grado de informática hace necesaria una adecuada integración de contenidos según lo estipulado por la Res. ME Nº 786/09. En ella se establece, en el ANEXO IV-1 -Estándares para la Acreditación de las carreras de Licenciatura en Ciencias de la Computación, Licenciatura en Sistemas de Información …, en lo que refiere a:

I. Contexto institucional

...

I.3. La institución debe tener definidas y desarrollar políticas institucionales en los siguientes campos:
a) investigación científica básica y aplicada.
b) desarrollo tecnológico y transferencia.
c) actualización y perfeccionamiento del personal docente y de apoyo, que no se limitará a la capacitación en el área científica o profesional específica y a los aspectos pedagógicos, sino que incluirá también el desarrollo de una adecuada formación interdisciplinaria.
...

II. Plan de estudios y formación

...

II.5. En el plan de estudios los contenidos deben integrarse horizontal y verticalmente. Asimismo deben existir mecanismos para la integración de docentes en experiencias educacionales comunes.
...

II.7. El plan de estudios debe incluir formación experimental de laboratorio, taller y/o campo que capacite al estudiante en la especialidad a la que se refiera el programa.

II.8. El plan de estudios debe incluir actividades de resolución de problemas del mundo real (reales o hipotéticos) con utilización de fundamentos, metodologías e instrumentos informáticos, en las que se apliquen los conocimientos de la currícula.

II.9. El plan de estudios debe incluir actividades de proyecto y diseño de sistemas informáticos, contemplando una experiencia significativa que requiera la aplicación integrada de conceptos fundamentales de la currícula (Ciencias Básicas, Teoría de la Computación, Algoritmos y Lenguajes, Ingeniería de Software, Bases de Datos y Sistemas de Información, Arquitectura, Sistemas

Operativos y Redes, Aspectos Profesionales y Sociales), así como habilidades que estimulen la capacidad de análisis, de síntesis y el espíritu crítico del estudiante, despierten su vocación por la innovación y entrenen para el trabajo en equipo y la valoración de alternativas.
…

Procedimiento

Del análisis y revisión de los planes de estudio correspondientes a las carreras de informática, los tópicos y técnicas abordadas en distintas asignaturas, además de la bibliografía pertinente, se encuentra "en teoría" una estrecha vinculación entre el proceso educativo y las TIC. Esta concatenación, muchas veces no se ve reflejada en la práctica. Si bien existe una vinculación directa entre las asignaturas de una misma área, ésta no se plantea de igual forma entre las de distintas áreas de conocimiento. Por ello, en este trabajo sólo se hará referencia a un caso correspondiente a esta última cuestión. Dentro de las áreas: Algoritmos y Lenguajes, Ciencias Básicas, Ingeniería De Software – Base de Datos – Sistemas de Información y Teoría de la Computación, se consideran las asignaturas Algoritmos y Estructuras de Datos, Sistemas de Datos, Probabilidad y Estadística, Base de Datos, Base de Datos Avanzadas y finalmente a Inteligencia Artificial, pues son las que aportan contenidos sustanciales a la MD.

Tomando a la MD como broche final en donde concurren conceptos, técnicas, métodos y herramientas impartidas a lo largo de la currícula y que genera un espacio propicio para el ejercicio profesional a través de su aplicación a problemas del mundo real, se procura que el futuro licenciado se desempeñe exitosamente.

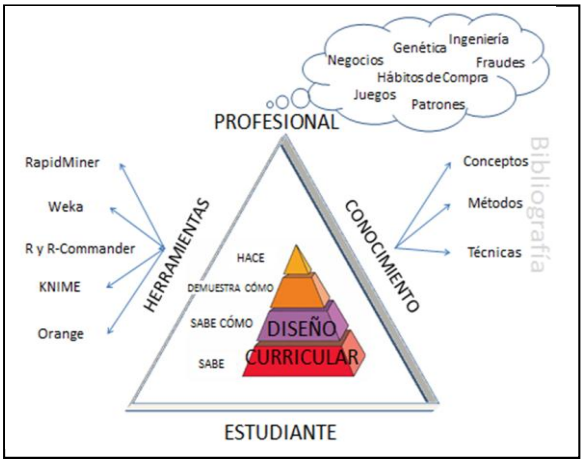

Figura 1-Etapas y componentes en un proceso de formación profesional.

En este contexto, la MD toma conceptos impartidos en Estadística tales como Estadística Descriptiva, Probabilidad, Análisis de Varianza, Regresión, Pruebas Chi-cuadrado, entre otros. A la vez tiene inmersos temas de Base de Datos - Datawarehouse-, de Aprendizaje de Máquina- Redes Neuronales, Algoritmos Genéticos, Inteligencia de Enjambres- y Estrategias de Diseño de Algoritmos dictados en las asignaturas anteriormente mencionadas, entre otros aportes de la Informática.

Es importante a partir de este análisis tener en cuenta, para el proceso de enseñanza aprendizaje, las siguientes cuestiones:

- Diferenciación entre las distintas técnicas contenidas en el desarrollo curricular.
- Compatibilidad en el uso de la terminología.
- Utilización de bibliografía coincidente.
- Disponibilidad de herramientas computacionales integrables

Minería de Datos versus Estadística

En MD y en Estadística se utilizan técnicas similares para resolver problemas semejantes,

pero sus enfoques difieren en diferentes aspectos:

- Las técnicas estadísticas son en su mayoría técnicas confirmatorias, mientras que las técnicas de MD son generalmente exploratorias. Por ello cuando no existen supuestos de partida y se busca conocimiento que proporcione información novedosa para la toma de decisiones, se puede aplicar MD.
- Tradicionalmente se realizan estadísticas sobre conjuntos de datos más bien pequeños (de unos pocos miles de registros o filas), con una reducida cantidad de atributos (pocas columnas); cuantas más variables entran en el problema, más difícil resulta encontrar hipótesis de partida interesantes. Por esta razón la teoría del muestreo, aplicada por ejemplo en la realización de encuestas, es un tópico importante de este área de conocimiento. Por otro lado, en MD se asume que existen grandes cantidades de datos disponibles y poder de procesamiento más que suficiente.
- Dadas las grandes dimensiones de los conjuntos de datos que pueden ser examinados, se debe confiar mucho en su procesamiento automático, por lo que la MD, a diferencia de la Estadística, hace mayor hincapié en los algoritmos. Es fácil desarrollar nuevos algoritmos, pero sin perder de vista la teoría subyacente para poder evaluar el progreso realizado.
- Las estadísticas originalmente derivan de medir magnitudes científicas. Estas mediciones son cuantitativas y el valor medido exacto depende de diversos factores, por lo que los valores observados son tratados por la estadística dentro de un intervalo de confianza. Por el contrario, en MD se tiende a ignorar los posibles errores de medición.
- La MD tiene que trabajar dentro de las limitaciones de las prácticas comerciales existentes, por lo que diseñar experimentos en el mundo de los negocios es una tarea dificultosa.
- Si los datos tienen poca variación en el tiempo se justifica la inversión en un análisis estadístico pues los resultados que se obtengan perdurarán. Por otro lado, si los datos varían dinámicamente, las técnicas de MD permiten explorar los cambios y determinar cuando se presentan modificaciones, lo que incide en decisiones a corto o mediano plazo.
- Cuando el objetivo de la investigación es encontrar causalidad es más adecuado utilizar técnicas de estadística dado que en los resultados que se obtienen de aplicar MD es dificultoso descubrir relaciones del tipo causa- efecto.
- Cuando se aspira obtener conclusiones extensibles a otros elementos de poblaciones similares, es conveniente el empleo de la inferencia estadística. Esto viene relacionado con situaciones en las que se dispone sólo de muestras, con el consiguiente problema de aportar validez a las mismas. Con MD se generan modelos que luego deben validarse con otros casos conocidos de la población, utilizados a modo de testeo.

Puede decirse que MD y Estadística son complementarias, no excluyentes, y permiten obtener conocimiento inédito de los datos o dar respuestas a cuestiones concretas de negocio u otra problemática. Así en el desarrollo de un proyecto de MD se recurre a la estadística, por ejemplo al momento de preparar los datos (tratamiento de valores erróneos, valores omitidos,...), aproximación de las distribuciones de las variables en estudio y posible generación de hipótesis a refutar con una metodología o técnica estadística.

Terminología

Tanto la Estadística como la Inteligencia Artificial, entre otras disciplinas, contribuyen al desarrollo de la MD. Así, a la Estadística

Exploratoria y la Inteligencia Artificial se las asocia con el Análisis de Datos y a la Estadística Inferencial con pruebas de hipótesis.
La terminología de ambas disciplinas difiere, como ejemplo se puede citar el problema de predicción por redes neuronales –Figura 2– [2]

Inteligencia Artificial	***Estadística***
red (network)	modelo
ejemplos (patterns)	observaciones, individuos
features, inputs, outputs	variables
inputs	variables explicativas
outputs, targets	variables de respuesta
errores	residuos
training, learning	estimación
función de error	criterio de ajuste
pesos, coef. sinápticos	parámetros
aprendizaje supervisado	regresión, discriminación
aprendizaje no supervisado	clasificación

Figura 2 - Equivalencias de nomenclatura entre la Estadística y la Inteligencia Artificial para el problema de predicción por redes neuronales

Esta situación lleva a reflexionar respecto a una problemática existente, aunque no abordada como tal y que consiste en el uso de términos diferentes que refieren al mismo concepto. Este aspecto es tratado por Llavona Arregui [15] en su tesis doctoral "Terminología de Estadística y Minería de Datos en Lengua Inglesa". Este autor considera a la Estadística como la rama de las matemáticas que se ocupa de reunir y describir datos y a la MD como la extracción de información oculta, de valor predictivo –diferencia fundamental con la Estadística –, de grandes bases de datos. En este trabajo, en el que la MD se reveló como una ciencia independiente de la estadística, se seleccionó un corpus sobre MD conformado por las diversas temáticas propias de la materia, de una extensión aproximada de 200.000 palabras. Como resultado se confeccionó un glosario compuesto por 114 términos en Español-Inglés. Al analizar el origen de los términos, a partir de ubicar el concepto al que refiere con la mayor exactitud posible en su campo de conocimiento, se estableció que:

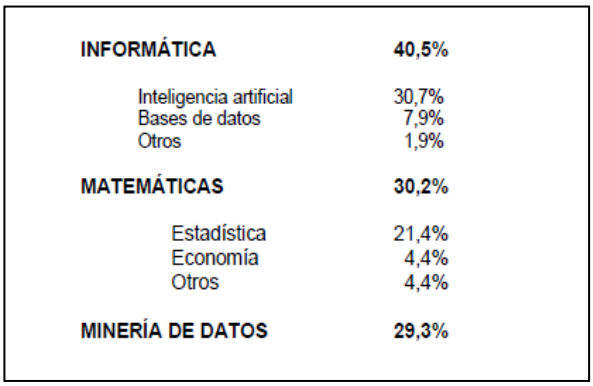

Figura 3 - Origen de los términos que constituyen el campo de la minería de datos

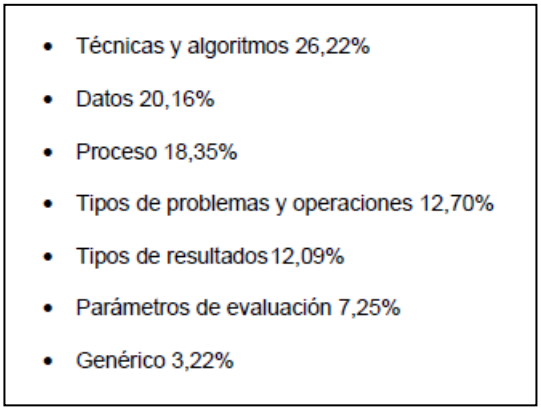

Figura 4 - Principales sub-apartados en minería de datos y porcentajes de términos que comprenden

Se puede observar a partir de la terminología empleada, Figuras 3 y 4, la importante vinculación que existe entre Minería de Datos, Inteligencia Artificial y Estadística. A la vez, al analizar el sub-apartado de minería de datos en el que se ubica el término, se encontró que el mayor porcentaje provenía del área Técnicas y algoritmos.

Este punto debería estar explícitamente presente en la currícula de una carrera en la que se impartan las temáticas mencionadas no solo a través de correlatividades entre las asignaturas que las contienen, sino también entre las estrategias empleadas para transmitirlas en el proceso de enseñanza aprendizaje. En otras palabras, esta conexión entre contenidos de materias pertenecientes a áreas posiblemente diferentes de una currícula, respaldan adecuadas y necesarias vinculaciones horizontales y verticales requeridas por el proceso de acreditación de carreras de Informática. En particular, en las carreras del Departamento de Informática de la FCEFN de la UNSJ, la asignatura Inteligencia Artificial pertenece al área "Teoría de la Computación", la asignatura Probabilidad y Estadística está en el área "Ciencias Básicas", mientras que los contenidos de MD y de Técnicas y algoritmos se imparten en las áreas "Ingeniería de Software-Base de Datos-Sistemas de Información" y "Algoritmos y Lenguajes", respectivamente.

Herramientas de computacionales

Numerosos estudios se han centrado en analizar la popularidad de los software en MD. Así se puede encontrar que, R es señalado en primer lugar, siendo R un programa básicamente estadístico. Figura 5. [4]

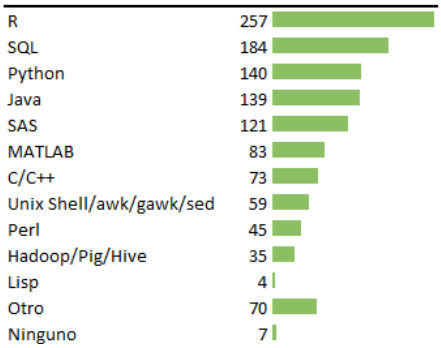

Figura 5 - Respuesta de 570 personas consultadas sobre ¿Qué herramientas de minería/análisis de datos utilizó

También, en una consulta respondida por 1103 votantes, se indica a RapidMiner como una de las herramientas de uso más frecuente en MD. Figura 6. [5]

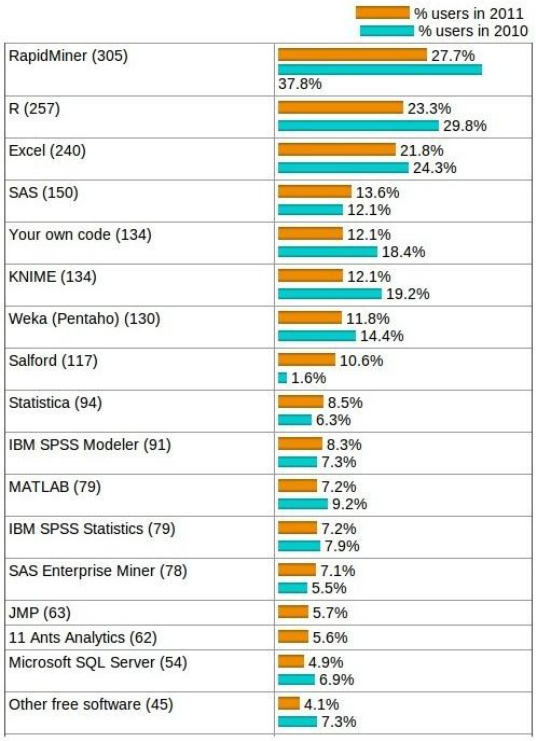

Figura 6 - Algunas de las respuestas de los 1103 votantes a la pregunta: ¿Qué herramientas de minería/análisis minería/análisis de datos que utilizó en los últimos 12 meses para un proyecto real (no sólo para evaluación)?

Dado que R también se puede tratar como una extensión de RapidMiner, lo adecuado es trabajarlos de manera integrada. Esto constituye un aporte tendiente a eliminar brechas que en diferentes oportunidades se producen entre áreas disciplinares y/o asignaturas que hacen uso de TIC en el proceso de enseñanza aprendizaje.

Entre las herramientas de software libre más usadas – Figura 5- orientadas al área de la minería de datos, minería de texto, minería web y utilizadas por la comunidad científica están, entre otras, RapidMiner (anteriormente, YALE, Yet Another Learning Environment de

la Universidad de Dortmund, Alemania), R (The R Project for Statistical Computing), KNIME (Konstanz Information Miner, de Universidad de Constanza Alemania), y Weka (Waikato Environment for Knowledge Analysis de la Universidad de Waikato Nueva Zelandia). Todas estas herramientas computacionales son bancos de pruebas de algoritmos aplicables a tareas de preprocesamiento, propuestas de clasificación, segmentación y reglas de asociación, como de distintas formas de visualización de resultados.

Cada una de estas herramientas, si bien tienen características propias que las distinguen y particularizan, poseen interfaces que las relacionan con los formatos y algoritmos existentes en cada una de las herramientas informáticas nombradas. Esto puede ser apreciado en el entorno de presentación de KNIME y RapidMiner, Figura 7, donde se observa la factibilidad de integración con las aplicaciones y extensiones R y Weka, dejando al usuario la elección desde su familiaridad con un entorno de software.

Figura 7 - Propuestas de enlaces con otras herramientas de software en KNIME 2.5.4 y RapidMiner 5.2.003

Cada una de estas herramientas, en permanente evolución, ofrece dos o tres actualizaciones por año. Con ellas y desde la práctica áulica los alumnos tendrían la posibilidad, entre otras de:

- Interconectar administradores de bases de datos y acceder a repositorios con diferentes formatos.
- Comprobar complejidades temporales y espaciales de distintos algoritmos. Comparar algoritmos, cuando se aplican a un mismo conjunto de datos.
- Utilizar y comparar analíticamente, diferentes alternativas de visualización de resultados.
- Diseñar e implementar, en el lenguaje asociado a la herramienta y en el caso de no contar con alguna métrica o estrategia algorítmica, nuevos algoritmos.

Así, las herramientas consideradas permiten al alumno, desde las diferentes asignaturas, un continuo crecimiento y profundización en el conocimiento de temas asociados e inherentes a la MD.

También, desde el material bibliográfico provisto por Ian Witten y Eibe Frank, "Data Mining - Practical Machine Learning Tools And Techniques, 2Nd Edition (2005)" se transita por el conocimiento del aprendizaje de máquina, desde la estadística a la MD interrelacionando conceptos teóricos con la práctica pertinente en el entorno de la herramienta de software Weka, desarrollada por los mismos autores.

DISCUSIÓN

A lo largo del trabajo se ha señalado la importancia que tiene la vinculación entre conceptos, técnicas, herramientas, terminología. Por ello se sugiere tener en cuenta los siguientes considerandos:

- Evidenciar las diferencias entre las técnicas para evitar la superposición de contenidos y distinguir cuando corresponde su aplicación.
- Usar la misma terminología o bien hacer alusión a las distintas formas en cada uno de los espacios curriculares que refieren al mismo concepto.
- Utilizar y recomendar bibliografía común.
- Aprovechar los beneficios del uso integrado, bajo un mismo entorno, de

herramientas computacionales y promover la adquisición de destrezas con herramientas de código libre y abierto.

- Conformar equipos de investigación interdisciplinarios de modo de lograr recursos humanos con una formación compuesta por las temáticas que abordan.

En resumen, para lograr la integración a lo largo del desarrollo curricular de las carreras universitarias, es necesario abordar tanto la disyunción entre teoría y práctica de cada asignatura como con las restantes. Ello podría contribuir a superar la desvinculación de la formación superior con la práctica profesional y de esta manera construir puentes que las acerquen.

Referencias

[1] Adams, Niall M.; Blunt, Gordon; *Hand*, David J.; Kelly Mark G. *Data Mining for Fun and Profit*. Source: Statist. Sci. Volume 15, Number 2 111-131. 2000

[2] Aluja, T *La minería de datos, entre la estadística y la inteligencia artificial*. QÜESTIIÓ, vol 25,3, p 479-498- http://www.idescat.cat/sort/questiio/questiiopdf/25.3.4.Aluja.pdf. 2001

[3] Berry, Michael J.A.; Linoff, Gordon S. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management - Second Edition*- Wiley Publishing. 2004

[4] Figura 5. http://www.r-bloggers.com/kdnuggest-r-most-commonly-used-software-for-data-mining-analytics/

[5] Figura 6. http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html

[6] Giudici P. *Applied Data Mining- Statistical Methods for Business and Industry*. Wiley. 2003

[7] Han, J; Kamber, M. *Data Mining: Concepts y Techniques***.** *Second Editions.* Morgan Kaufmann Publisher. 2006.

[8] Hand, David; Mannila, Heikki; Smyth, Padhraic *Principles of Data Mining.*The MIT Press. 2001

[9] Hastie ,T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[10] Hernández Orallo J., Ramirez Quintana, J, Ferri Ramirez, C. *Introducción a la Minería de Datos*. Pearson-Prentice Hall. 2008

[11] Larose, D. *Data Mining. Methods and Models.* Department of Mathematical Sciences Central Connecticut State University Wiley. A John Wiley & Sons, Inc Publication. 2006.

[12] Larose, D. *Discovering Knowledge In Data - An Introduction to Data Mining.* John Wiley & Sons, Inc., Publication. 2005

[13] Llavona Arregui, José Luis () *Terminología de estadística y minería de datos en lengua inglesa.* Tesis Doctoral. http://eprints.ucm.es/11032/. 2010

[14] Myatt, G.J. *Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining*. Wiley-Interscience a John Wiley & Sons, Inc., Publication. 2007

[15] R http://www.r-project.org/

[16] Rapid-I http://rapid-i.com/api/rapidminer-5.1/com/rapidminer/tools. 2011

[17] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin *Introduction to Data Mining* Addison-Wesley. 2006

[18] Torgo, Luis. *Data Mining with R: Learning with Case Studies*. Chapman and Hall/CRC. 2010

[19] Unwin, A.; Theus, M.; Hofmann, H. *Graphics of Large Datasets: Visualizing a Million (Statistics and Computing)*. Springer, 2006

[20] Web 1- http://www-users.cs.umn.edu/~kumar/dmbook/resources.htm

[21] Web 2 - http://www.junauza.com/2010/11/free-data-mining-software.html