UNA PROPUESTA DE INTEGRACIÓN DE INTERFACES DE USUARIO EN MÉTODOS DE MINERÍA DE DATOS

SONIA I. MARIÑO - PEDRO L. ALFONZO

Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura.
Universidad Nacional del Nordeste
simarinio@yahoo.com-plalfonzo@hotmail.com

Fechas recepción: Diciembre 2016 - Fecha aprobación: Abril 2017

RESUMEN

La Minería de Datos ofrece una diversidad de algoritmos estadísticos e inteligentes para transformar los datos en información, y deriva en conocimiento que apoya la toma de decisiones. Actualmente, existe una diversidad de herramientas software que los implementan y requieren de un conocimiento adicional para su adecuada utilización. Ante esta situación se propone un modelo de proceso que integra un método de Minería de Datos y el diseño de interfaces para facilitar a los potenciales usuarios el empleo de los modelos de conocimiento. A efectos de la validación se seleccionó un dominio de la botánica.

PALABRAS CLAVE: Minería de Datos - Herramientas - Modelos de Proceso - Interfaces de Usuario.

ABSTRACT

Data Mining offers a diversity of statistical and intelligent algorithms to transform data into valuable information, and to produce knowledge oriented to support decision making. Also, there are diversity of software tools that implement algorithms, so in this order, is required an additional knowledge to ensure proper use. The paper presents a process model, integrating Data Mining method and design of interfaces, to ensure the users application. For validation purposes, the model is applied in a botanical domain.

KEYWORDS: Data Mining - Tools - Process Models - User Interfaces.

1. INTRODUCCIÓN

El incremento de las tecnologías de la información y aquellas dirigidas al descubrimiento de conocimiento en conjuntos de datos ha permitido el avance del procesamiento de la información dirigida a la toma de decisiones.

En el año 1996 surge el término *Knowledge Discovery in Databases* (KDD), o modelo descubrimiento del conocimiento. Éste establece las etapas principales de un proyecto de explotación de información (Frawley, Piatetsky-Shapiro, Matheus, 1992; Matheus. Chan, Piatetsky-Shapiro, 1993). Moine,

Haedo, Gordillo (2011a, 2011b), especifica a la Minería de Datos (MD) como la etapa del proceso en la cual se realiza la extracción de patrones a partir de los datos.

Mientras que el KDD trata el proceso necesario para generar conocimiento a partir de los datos, la MD se entiende como la actividad de aplicar distintos algoritmos en los datos para obtener patrones. La MD proporciona las herramientas para automatizar el proceso de análisis de datos y el artesanal proceso estadístico de selección de hipótesis. Es decir, se entiende a la Minería de Datos como un subproceso que se integra al proceso general destinado a obtener patrones de conocimiento (Martins, Pesado, García Martínez, 2014).

Moine *et al.* (2011a, 2011b,) indican que en la actualidad "el término KDD y Minería de Datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento". Por lo expuesto, el descubrimiento del conocimiento y la Minería de Datos contribuyen a la toma de decisiones tácticas y estratégicas, proporcionando un sentido automatizado para la generación de conocimiento y por ende su aplicación es amplia en las diferentes ramas de la investigación (Valcárcel Asencios, 2004).

El KDD, comprende diferentes métodos como los cuantitativos, los probabilísticos y los estadísticos (Valcárcel Asencios, 2004). Su utilización depende del comportamiento de los algoritmos para transformar datos en información y ésta en conocimiento.

Ejecutar un proceso de KDD implica adoptar un método o modelo de proceso de Minería de Datos, como por ejemplo CRISP-DM (Chapman *et al*., 2000).

Para el procesamiento de los datos, el método de Minería de Datos recurre a alguna herramienta disponible e integradora de la diversidad de estos algoritmos. Entre aquellas de libre distribución se mencionan Weka, Rapidminer, Tanagra.

Las herramientas de MD, proporcionan una diversidad de algoritmos para el procesamiento de los datos bajo esquemas supervisados o no supervisados, facilitando a los usuarios la especificación de los valores de los parámetros según los distintos modelos inferenciales.

Un aspecto a considerar es el grado de comprensión de estos artefactos de las Tecnologías de la Información (TI) por aquellos especialistas en un dominio interesados en obtener resultados para la toma de decisiones sin profundizar en el manejo de las herramientas computacionales.

Sin embargo, ¿son realmente estos modelos de conocimiento construidos con herramientas de MD utilizados por los destinatarios? Ante esta cuestión se elaboró la propuesta que a continuación se describe.

Así se define como objetivo del trabajo: exponer un modelo de proceso software que integra la Minería de Datos y la implementación de interfaces para facilitar procesos decisorios de los potenciales usuarios quienes interactúan con modelos de conocimiento previamente entrenados.

2. METODOLOGÍA

El método general abordado en el trabajo retoma el concepto de fases (Samaja, 2003). Es decir, se entiende como una sucesión de momentos que pueden solaparse, consistiendo en:

Fase 1: Revisión sistemática de la literatura referente a los métodos de Minería de Datos. Se optó por aquellos comprendidos en el aprendizaje supervisado, eligiéndose métodos bayesianos (Jensen, 1996; 2001; Castillo, Gutiérrez y Hadi, 1997; Sucar, 2006; Rusell y Norving, 2004; Moret Bonillo, 2014).

Fase 2: Revisión sistemática de la literatura referente a herramientas de Minería de Datos, entre las que se mencionan Weka, Rapidminer, Tanagra. En este trabajo se eligió Weka, sus siglas significan Entorno Waikato para el Análisis del Conocimiento, y que:

- Implementa tareas de Minería de Datos, especialmente pre procesamiento de datos, agrupamiento, clasificación, regresión, visualización y selección.
- Aplica algoritmos sobre la hipótesis que los datos están disponibles en un único archivo plano o relación. Cada fila del archivo representa un patrón que el algoritmo debe aprender para aplicar el modelo inferencial en procesos decisorios.
- Posibilita la construcción de interfaces para usuarios finales en el lenguaje de programación Java y permite modificaciones por su característica de código abierto.

Fase 3: Diseño de un modelo de proceso para la generación de artefactos inteligentes, focalizado en el desarrollo de interfaces de usuarios para el acceso a modelos entrenados en herramientas de Minería de Datos y el despliegue de los resultados. En trabajo se basó en CRISP-DM. Dado que el trabajo se centra en el modelo de proceso, éste se describe en la sección 3.

3. RESULTADOS

A continuación se explicita la propuesta de modelo de proceso para la generación de artefactos inteligentes integrando interfaces de usuario y modelos de conocimiento previamente entrenados y validados con una herramienta de MD, y se incluye su validación en un dominio específico.

3.1 Modelo proceso propuesto

A continuación se describe el modelo de proceso que consta de 3 (tres) fases y está representado en la Figura 1. La primera sección de la Figura ilustra la Fase 1 del modelo; en el centro se indica la Fase 2, principal innovación de la propuesta sustentada en integrar el proceso de MD desde interfaces inteligibles para los potenciales destinatarios. La sección inferior representa la Validación de la propuesta o Fase 3.

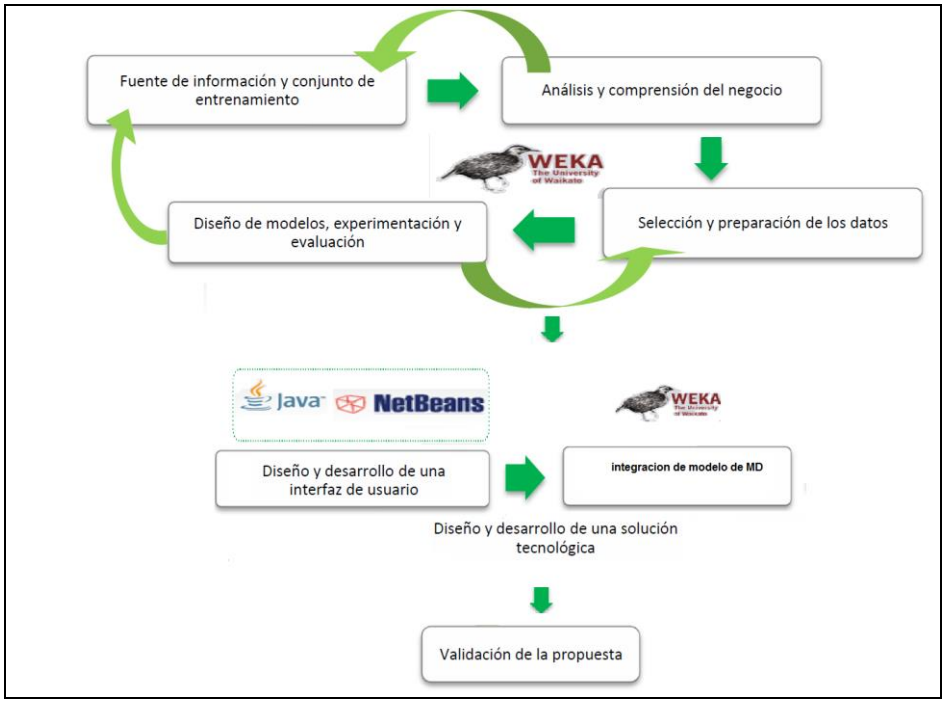

FIGURA 1. Modelo de proceso propuesto (Fuente elaboración propia)

Fase 1. Modelo de proceso basado en CRISP-DM: Se elige CRIPS-DM como modelo de proceso de la Minería de Datos para construir los modelos de conocimiento de un dominio. Cabe aclarar que lo expuesto en esta fase se describió en Mariño y Alfonzo (2016).

- **Fuente de información y conjunto de entrenamiento**

Se selecciona la fuente de información, de la cual se derivan los datos para la construcción del modelo de conocimiento. Se determinan como variables evidenciales los caracteres seleccionados por el EDC (Experto en el Dominio de Conocimiento). Se define la variable objetivo o meta.

- **Análisis y comprensión del negocio**

Se comprenden los objetivos y requerimientos del proyecto desde una perspectiva del dominio que faciliten la definición de un problema posible de resolver con MD.

- **Selección y preparación de los datos**

Preparación de datos: Involucra aquellas actividades que facilitan construir el conjunto de datos final, el cual se utiliza como entrada en las

herramientas de modelado. Las tareas se pueden aplicar múltiples veces, y sin un orden pre-establecido. Incluyen procesos de extracción, transformación y carga, proceso conocido como ETL (Simitsis y Vassiliadis, 2008).

Transformación de los datos en formato legible por la herramienta elegida de MD. Se preparan los datos según el formato requerido por el software de modelado y simulación computacional, en este caso Weka.

Estimación de estadísticos sobre los atributos: Registrados los datos, desde la herramienta de MD se determinan los atributos y se computan algunas estadísticas básicas sobre cada atributo. Si el conjunto de datos seleccionados son atributos continuos/numéricos, se visualizan valores mínimo, máximo, media, desviación estándar, entre otros.

- **Modelado**

Se seleccionan y aplican distintas técnicas de modelado según el conjunto de datos disponibles y la finalidad de la simulación. Se requiere disponer de numerosos ejemplos para lograr aprendizajes significativos. Es decir, pocos ejemplos dificultan la inferencia de los rasgos o evidencias que distinguen los distintos valores que asumen la variable objetivo e incide en los resultados generados.

Se trabaja con un especialista del dominio, quien selecciona los patrones de datos y las variables evidenciales implicadas, y valida los resultados generados por el modelo de conocimiento.

- **Diseño de modelos, experimentación y evaluación**

Se define la tarea de Minería de Datos para proponer una respuesta al problema planteado:

- Se elige el algoritmo de entrenamiento del modelo.
- Si corresponde, se seleccionan los estimadores asociados al modelo inferencial.
- Se diseñan diversos experimentos, modificándose en cada uno de ellos, los estimadores.
- Se procede a la simulación, interpretación y valoración de los resultados, confrontándose con aquellos valores meta esperados y determinados por el especialista del dominio.
- Se determinan las métricas o medidas de calidad. Para evaluar la efectividad del proceso inferencial, se puede optar por las siguientes métricas de precisión: el estadístico Kappa, el Error Medio Absoluto (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el parámetro *Alpha* que indica el valor de la tasa de aprendizaje.
    - Kappa *Statistic*, medida de concordancia entre las categorías pronosticadas por el clasificador y las categorías observadas, considera las posibles concordancias debidas al azar. La valoración del índice Kappa está dada:
        - Si el valor es 1: Concordancia perfecta.

- Si el valor es 0: Concordancia debida al azar.
- Si el valor es negativo: Concordancia menor que la esperada por azar.

  - RMSE o *Root Mean Squared Error*, mide las diferencias entre los valores brindados por un modelo o un estimador y los valores realmente observados. Es una medida de precisión, permite comparar diferentes errores de predicción de un mismo conjunto de datos, dado que es dependiente de la escala muestra.
  - MAE o *Mean Absolute Error*, se define como el Error Absoluto de una Media a la diferencia entre el valor medio obtenido y el hallado en esa media. El promedio de error absoluto, es la suma de los errores absolutos de clasificación en cada uno de los sujetos llevados a promedio. El clasificador que arroje mayor cifra (mayor a 0.1) define un error de clasificación alto, por lo cual no se debe considerar sobre aquellos que arrojen una cifra menor.

- Parámetros de selección de los modelos, se determinan los valores de los indicadores para elegir los modelos, por ejemplo: Clasificación correcta > 90 % instancias; Clasificación incorrecta < 10 % instancias; MAE <0.1 ideal; Kappa *statistic* >0.79, >0.9 ideal; RMSE < 0.3, <1 ideal. El clasificador que cumpla con estas especificaciones, se considera representativo para su integración en un sistema inteligente que emulará al especialista del dominio ante nuevos casos de identificación. Si los modelos proporcionan un mismo valor en la métrica MAE, se considerará aquel menor valor de RMSE, para seleccionar el modelo más representativo.
- Comprobación, entrenado el conjunto de datos, se procede a su verificación utilizando la técnica denominada *Percent Split* en la herramienta de MD seleccionada. Esta opción divide los datos en dos grupos, el porcentaje especificado representa las instancias utilizadas para construir el modelo, y éste es evaluado respecto a las restantes. Cuando el número de instancias es suficientemente elevado, esta opción es suficiente para estimar con precisión las prestaciones del clasificador en el dominio.

Fase 2. Diseño y desarrollo de una solución tecnológica

Se elabora una solución tecnológica considerando los sujetos o actores (Figura 2), quienes intervienen en los distintos contextos y en el proceso decisorio a través de una tecnología inteligente.

A cada perfil del modelo de proceso software se asigna un conjunto de privilegios y permisos que definen las funcionalidades asignadas en los procesos de entrenamiento y de utilización, identificándose: Ingeniero de

Sistemas (ISist), Ingeniero del Conocimiento (IC), Experto en el Dominio de Conocimiento (EDC), Usuarios Finales (UF) y establecidos en Mariño (2015). En el trabajo se entiende como: i) Ingeniero de Sistemas al perfil representado por el diseñador, desarrollador y otros especialistas de las Ciencias de la Computación; ii) Experto en el Dominio de Conocimiento, quienes brindan la información a representar y simular; iii) Usuarios Finales, los especialistas en el dominio de conocimiento, la comunidad, otros especialistas, los aprendices y/o principiantes.

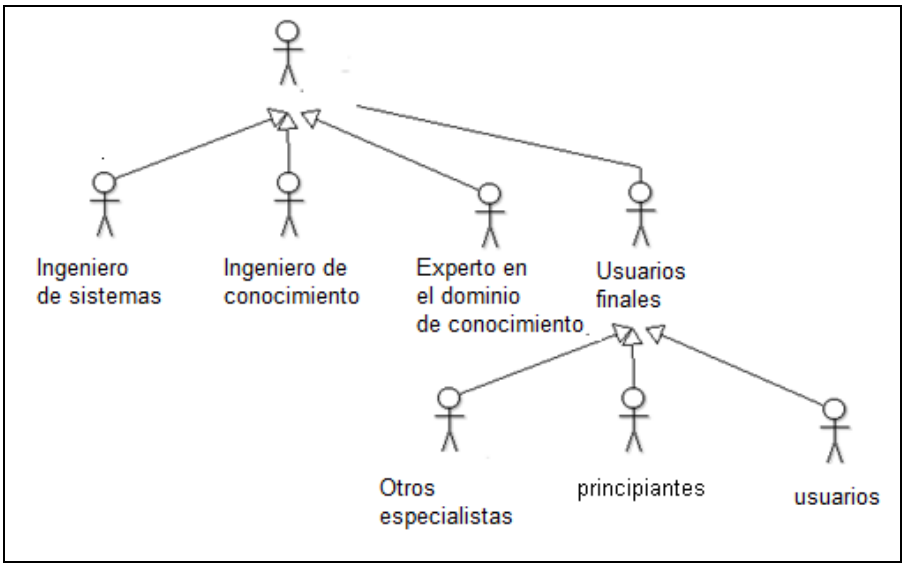

Figura 2. Perfiles de sujetos intervinientes en el modelo (Fuente: Mariño, 2015)

La elaboración del software inteligente implica las siguientes actividades:

- Selección de lenguajes de programación que facilitan la interacción de los algoritmos de MD desde una interfaz de usuario. Particularmente, en esta propuesta se definen como requerimientos de conocimiento para el ISist:
  - ✓ una herramienta de Minería de Datos como Weka,
  - ✓ una infraestructura *front-end*, como *NetBeans*,
  - ✓ el diseño y desarrollo de aplicaciones en código Java. Este requerimiento permite acceder a los algoritmos de aprendizaje de maquina codificados en la herramienta Weka a través de clases de Java, y presentar los resultados en la interfaz de usuario diseñada.

- Diseño y desarrollo de una interfaz de usuario
  Para acceder a procesos decisorios basados en técnicas de MD, y que permitan a los usuarios interactuar con modelos de conocimiento previamente entrenados se diseña y desarrolla una interfaz de usuario. En

el caso de estudio, ésta mediará en procesos de identificación botánica, según se ilustra en el punto 3.2.

Fase 3: Validación de la propuesta

A fines de la validación y con miras a lograr el uso de estos esquemas de representación del conocimiento y apoyo a la toma de decisiones dirigida a los Expertos en un Dominio y los Usuarios Finales, se diseña un caso de estudio.

Cabe aclarar que, dado que estos modelos representan y tratan el conocimiento de un dominio, por ello, la interfaz representa los objetos tratados en la problemática elegida.

3.2 Caso de estudio

En Botánica el proceso de clasificación computacional se asocia a la identificación de taxones, numerosos expertos de este dominio desconocen el uso de herramientas de Minería de Datos. Por lo expuesto, se entiende que esta propuesta aportaría en el proceso decisorio basado en tecnologías inteligentes accesibles desde interfaces de usuario destinadas a especialistas en dominios de conocimientos.

En este caso, se integran modelos bayesianos (Jensen, 1996; 2001; Castillo *et al*., 1997; Sucar, 2006; Rusell y Norving, 2004; Moret Bonillo, 2014), siendo el algoritmo de entrenamiento seleccionado el denominado *BayesNet* y descripto en Mariño y Alfonzo (2016).

La Figura 3 representa la interfaz, desde la cual el usuario identificado como EDC o UF interactúa ante un nuevo proceso decisorio. El EDC o UF puede seleccionar los valores de las variables evidenciales observadas en el nuevo caso, al presionar el botón IDENTIFICAR el software ejecuta el modelo de razonamiento –previamente seleccionado- y proporciona la mejor respuesta basada en valores probabilísticos simulando el proceder de un especialista del dominio.

Desde la interfaz de usuario, el EDC puede validar el correcto funcionamiento del modelo que representa el razonamiento del experto del dominio. Es decir, aun cuando se carece de una funcionalidad que argumenta explícitamente la decisión proporcionada, las elecciones de los valores de las variables evidenciales realizadas en la interfaz sostienen los resultados generados.

FIGURA 3. Interfaz de procesos decisorios (Fuente elaboración propia)

4. CONCLUSIONES

En la actualidad existe una vasta literatura en torno a tecnologías de Minería de Datos y su aplicación en la solución de problemas de diversos dominios del conocimiento. Se localizaron numerosos trabajos que exponen la visualización de la información y el descubrimiento de conocimiento en patrones de datos, y que principalmente describen la lectura y análisis de los mismos.

Sin embargo, aún en el siglo XXI, en la sociedad del conocimiento que se transita existen usuarios de distintos dominios del conocimiento quienes podrían desconocer en profundidad como utilizar los modelos de conocimiento desde herramientas computacionales. Por ello, este trabajo se focaliza en la construcción de un modelo de proceso que integra desde interfaces de usuario el acceso a modelos de conocimiento construidos según CRISP-DM. La propuesta se sintetizó en la Figura 3.

Dado que los sistemas inteligentes se validan en contextos específicos, se optó por integrar a la interfaz de usuario modelos de conocimiento sustentados en algoritmos de redes bayesianas aplicadas a dominios botánicos y presentados en Mariño y Alfonzo (2016). Se concretó una validación interna consistente en probar la comprensión de la interfaz con usuarios de perfil informático y el correcto despliegue de la información generada.

Como líneas de trabajo futuras, se menciona la validación del artefacto inteligente desde la Experiencia de Usuario en el modelo de proceso, a fin de retroalimentar hacia fases anteriores y cumplimentar las expectativas y las necesidades de los expertos y de los potenciales usuarios de un dominio.

Desde una perspectiva tecnológica, con miras de mejorar su adopción en contextos específicos, se continúa trabajando en la inclusión de las siguientes funcionalidades:

- Despliegue de explicación de la propuesta de solución del sistema inteligente, finalizado el proceso de identificación
- Actualización permanente de la base de conocimiento, permitiendo que la información generada puede eliminarse tras una consulta o almacenarse como datos, que serán utilizados en futuros procesos inferenciales.
- Generación de información adicional relacionada a la variable objetivo o meta propuesta por el sistema de decisión, en esta caso complementando aspectos de la especie inferida por el modelo de razonamiento. Lo expuesto es viable dado que WEKA proporciona acceso a bases de datos SQL utilizando conectividad Java, y procesa el resultado como una consulta de base de datos.

5. REFERENCIAS

CASTILLO, E.; GUTIÉRREZ, J.M.; HADI, A.S. (1998): "SISTEMAS EXPERTOS Y MODELOS DE REDES PROBABILÍSTICAS". Ed. Academia Española de Ingeniería, España.

CHAPMAN, P.; CLINTON, J.; KEBER, R.; KHABAZA, T.; REINARTZ, T., SHEARER, C.; WIRTH, R. (2000). CRISP-DM 1.0 Step by step BIguide. Edited by SPSS. Documento en línea. Disponible en: http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. (1992): "KNOWLEDGE DISCOVERY IN DATABASES: AN OVERVIEW". AI Magazine Vol. 13, num. 3, 57-70. Documento en línea. Disponible en: http://aaaipress.org/ojs/index.php/aimagazine/article/viewFile/1011/929.

JENSEN, F. V. (1996): "AN INTRODUCTION TO BAYESIAN NETWORKS". London, UCL Press.

JENSEN, F. V. (2001): "BAYESIAN NETWORKS AND DECISION GRAPHS". New York, Springer.

MARIÑO, S. I. (2015), Inédito.

MARIÑO, S. I.; ALFONZO, P. L. (2016): "SIMULACIÓN DEL RAZONAMIENTO EN EL PROCESO DE IDENTIFICACIÓN BOTÁNICA BASADO EN REDES BAYESIANAS". Revista de la Escuela de Perfeccionamiento en Investigación Operativa, 24(39): 55-72, ISSN 1853-977.

MARTINS, S.; PESADO, P.; GARCÍA-MARTÍNEZ, R. (2014): "PROPUESTA DE MODELO DE PROCESOS PARA UNA INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN: MOPROPEI". Revista Latinoamericana de Ingeniería de Software, 2(5): 313-332, ISSN 2314-264.

MATHEUS J. C.; CHAN, P. K.; PIATETSKY-SHAPIRO, G. (1993): "SYSTEMS FOR KNOWLEDGE DISCOVERY IN DATABASE". IEEE, TKDE, special issue on Learning & Discovery in Knowledge-Based Databases, 1-16, Documento en línea. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.169&rep=rep1&type=pdf.

MOINE, J. M.; HAEDO, S. A.; GORDILLO, S. (2011a): "ESTUDIO COMPARATIVO DE METODOLOGÍAS PARA MINERÍA DE DATOS". XIII Workshop de Investigadores en Ciencias de la Computación p. 278-281, ISBN: 978-950-673-892-1.

MOINE, J. M.; HAEDO, S. A.; GORDILLO, S. (2011b): "ANÁLISIS COMPARATIVO DE METODOLOGÍAS PARA LA GESTIÓN DE PROYECTOS DE MINERÍA DE DATOS". VIII Workshop Bases de Datos y Minería de Datos (WBDDM), CACIC 2011 - XVII Congreso Argentino de Ciencias de la Computación, octubre 2011: p. 931-938

MORET BONILLO, V. (2014): "REPRESENTACIÓN DEL CONOCIMIENTO Y RAZONAMIENTO AUTOMÁTICO". Departamento de Computación. Facultad de Informática. Universidad de A Coruña.

RAPIDMINER: DATA SCIENCE, MACHINE LEARNING. Disponible en: https://rapidminer.com/

RUSSELL, S.; NORVIG, P. (2004): "INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO". 2da edición, Prentice–Hall Hispanoamericana.

SAMAJA, J. (2003): "EPISTEMOLOGÍA Y METODOLOGÍA. ELEMENTOS PARA UNA TEORÍA DE LA INVESTIGACIÓN CIENTÍFICA". Ed. EUDEBA.

SIMITSIS, A.; VASSILIADIS, P. (2008): "A METHOD FOR THE MAPPING OF CONCEPTUAL DESIGNS TO LOGICAL BLUEPRINTS FOR ETL PROCESSES", Decision Support Systems, 45(1), 22-40.

SUCAR, L. E. (2006): "REDES BAYESIANAS. BS ARAUJO, APRENDIZAJE AUTOMÁTICO: CONCEPTOS BÁSICOS Y AVANZADOS". Pearson Educación. TANAGRA, Disponible en: http://eric.univ-lyon2.fr/~ricco/tanagra/

VALCÁRCEL ASENCIOS, V. (2004): “DATA MINING Y EL DESCUBRIMIENTO DE CONOCIMIENTO”. Revista de la Facultad de Ingeniería Industrial. Vol. (7) 2: pp. 83-86 (2004) UNMSM ISSN: 1560-9146. Diciembre de 2004.

WEKA, Disponible en: http://www.cs.waikato.ac.nz/ml/weka/