Hacia un nuevo proceso de minería de datos centrado en el usuario

Aldair Antonio Aquino

Maestría en Sistemas Interactivos Centrados en el Usuario, Universidad Veracruzana

aquinoaldair@hotmail.com

Guillermo Molero-Castillo

Cátedras CONACyT, Facultad de Estadística e Informática, Universidad Veracruzana

ggmoleroca@conacyt.mx

Rafael Rojano

Facultad de Estadística e Informática, Universidad Veracruzana

rrojano@uv.mx

Resumen

El diseño centrado en el usuario es un concepto que ha ganado popularidad en los últimos años como un factor de calidad para el desarrollo de proyectos de software. La cohesión del diseño centrado en el usuario y la minería de datos aportan un nuevo enfoque metodológico con el objetivo de mejorar la interacción entre el usuario y el descubrimiento de conocimiento en volúmenes de datos. La principal aportación de esta propuesta es el diseño de un marco metodológico centrado en el usuario para el desarrollo de proyectos de minería de datos, asociando para esto la norma ISO 9241-210:2010 (Human-centred design for interactive systems) y el proceso CRISP-DM (Cross Industry Standard Process for Data Mining).

Palabras Clave: CRISP-DM, diseño centrado en el usuario, ISO 9241-210:2010, minería de datos, reconocimiento de patrones.

Abstract

User-centered design is a concept that has gained popularity in recent years as a quality factor for the development of software projects. The cohesion of user-centered design and data mining bring a new methodological approach in order to improve the interaction between the user and knowledge discovery in data volumes. The main contribution of this proposal is to design a methodological framework for user-centered development of data mining projects, for this case is used ISO 9241-210: 2010 (Human-centered design for interactive systems) and process CRISP-DM (Cross Industry Standard Process for Data Mining).

Keyboards: *CRISP-DM, design centered user, data mining, ISO 9241-210:2010, patterns recognition.*

1. Introducción

En la actualidad, el crecimiento exponencial de la información generada por los usuarios en los diversos campos de conocimiento ha suscitado un reto continuo para el análisis, descubrimiento y entendimiento de los datos, más allá de recurrir a métodos tradicionales, como consultas a bases de datos. Es por esto que surgen tecnologías especializadas como la minería de datos que permiten realizar estas actividades a través de la integración de diversas disciplinas como estadística, inteligencia artificial, bases de datos, aprendizaje automático, entre otras. En este sentido, la minería de datos es una disciplina de la ciencia de la computación que es considerada por el Instituto Tecnológico de Massachusetts (MIT, 2001) como "*una de las diez tecnologías emergentes más importantes del siglo 21, que cambiará el sentido de investigación en el mundo*". Sin embargo, para realizar descubrimientos de conocimiento en volúmenes de datos es necesario contar con un marco de trabajo que permita planificar y guiar el

proceso de desarrollo del proyecto. Actualmente las metodologías más conocidas para el desarrollo de proyectos de minería de datos son KDD (Knowledge Discovery in Databases), CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, Assess), Catalyst, Six Sigma, entre otras. Estas metodologías cumplen con el propósito principal de encaminar el desarrollo de un proyecto de minería de datos; sin embargo, ninguna de éstas consideran al usuario como factor importante en cada una de las etapas, es decir, tienen como prioridad efectuar las actividades desde un enfoque de resultados. Por lo que, surge la necesidad de disponer de un proceso de minería de datos centrado en el usuario, que sirva como marco de referencia, basado en etapas, para el desarrollo de proyectos de explotación de datos con el fin de alcanzar los objetivos de manera satisfactoria.

2. Minería de datos

La minería de datos es un dominio de la ciencia de la computación que permite el análisis de grandes cantidades de datos para encontrar y extraer patrones significativos útiles para el proceso de la toma de decisiones. Dado su avance natural, existen variadas definiciones, por ejemplo para Larose *et al.* (2014) es el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas a través de grandes cantidades de datos, utilizando técnicas estadísticas, matemáticas y reconocimiento de patrones; mientras que para Govindarajan y Chandrasekaran (2011) es el uso de algoritmos para extraer la información y patrones derivados por el proceso de descubrimiento de conocimiento en bases de datos. Además, se considera a la minería de datos como un campo interdisciplinario que involucra a otras áreas como (Hernández *et al.*, 2004): estadística, matemática, bases de datos, aprendizaje automático, visualización, cómputo paralelo y distribuido, entre otras. Sin embargo, no todas la tareas asociadas con grandes volúmenes de datos son consideradas como parte de la minería de datos, por ejemplo para hacer una diferenciación Tan *et al.* (2006) mencionan algunas de éstas (Tabla 1).

Tabla 1. Diferenciación sobre algunas aplicaciones de minería de datos.

Actividad	Minería de datos	Observación
Dividir a los clientes de una empresa según su género	No	Corresponde a una consulta a la base de datos
Dividir a los clientes de una empresa según su rentabilidad	No	Es un cálculo contable. Sin embargo, la predicción de la rentabilidad de un nuevo cliente si sería minería de datos
Ordenar la base de datos de los estudiantes según el código de alumno	No	Es también una consulta a la base de datos
Predecir el precio futuro de las acciones de una empresa utilizando los registros históricos	Si	Se trata de crear un modelo que pueda predecir el valor continuo del precio de las acciones. Este es un ejemplo de minería de datos, conocido como modelo predictivo
Monitoreo de la frecuencia cardiaca de un paciente para detectar anomalías	Si	Se necesita de un modelo para el análisis del comportamiento la frecuencia cardíaca y se da una alarma cuando se produce un comportamiento inusual del corazón (normal y anormal)
Monitoreo de ondas sísmicas para actividades de terremoto	Si	Esta problemática está relacionada con la minería de datos, específicamente con la clasificación. Se necesita un modelo para analizar los diferentes tipos de comportamiento de onda sísmica

Fuente: Tan *et al*. (2006)

De acuerdo a la Tabla 1, la minería de datos va más allá de tareas sencillas, como consultas a bases de datos, sino que involucra el análisis de grades volúmenes de información con el objetivo de encontrar patrones significativos que sirvan de apoyo en el proceso de la toma de decisiones en las diversas áreas de conocimiento, como: medicina, educación, finanzas, mercadotecnia, entre otras. En ese sentido, para llevar a cabo proyectos de minería de datos, ésta se apoya en procedimientos que le sirven de guía para el proceso de análisis y explotación de datos.

3. Procesos tradicionales de minería de datos

Para llevar a cabo proyectos de minería de datos es necesario recurrir a procesos que permitan planificar y guiar el desarrollo del proyecto. Actualmente entre los más conocidos destacan KDD, CRISP-DM, SEMMA, Catalyst y otros; los cuales permiten estructurar el desarrollo de los proyectos en una serie de etapas relacionadas entre si.

3.1 KDD

Un término común en la minería de datos es el descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés), que viene a ser un proceso iterativo significativo que consta de una serie de fases para la generación de conocimiento y la toma de decisiones. Hernández *et al*. (2004) hacen una diferencia entre estos dos términos, KDD como un proceso que consta de una serie etapas y minería de datos como una etapa dentro de este proceso. Las fases de KDD son (Fayyad *et al*., 1996; Hernández et al., 2004; Molero y Céspedes, 2014):

1. Integración y recopilación. Consiste en establecer un entendimiento del dominio de la aplicación y de los conocimientos previos relevantes. En esta fase se determina también la selección de un conjunto de datos que pueden ser obtenidos de diferentes fuentes, sobre los cuales se realiza el descubrimiento.

2. Selección, limpieza y transformación. En esta etapa se seleccionan y preparan los datos que se van a minar. Sin embargo, existen factores como el ruido o valores atípicos que afectan la calidad de los datos, por lo que ante esta situación la limpieza es una de las tareas más importantes, puesto que permite la selección de la técnica que más se ajuste al problema a resolver.

3. Minería de datos. Es la fase más representativa, se determina qué tipo de tarea es la más apropiada, ya sea agrupamiento, reglas de asociación, correlación, clasificación, regresión, entre otras. Los resultados obtenidos dependen de las

fases anteriores, por lo que existe la posibilidad de regresar a los pasos previos para requerir nuevos datos o para redefinir la solución al problema planteado.

4. Evaluación e interpretación. Los patrones descubiertos deben cumplir con tres propiedades: precisión, comprensibles e interesantes. En esta fase se evalúan e interpretan los patrones obtenidos. Algunas validaciones pueden ser a través de índices de evaluación, validación cruzada, matrices de confusión, entre otras.

5. Difusión y uso. Como última fase, el conocimiento descubierto debe de ser incorporado en algún sistema o simplemente documentarlo para su difusión a las partes interesadas. Este proceso incluye también la revisión y resolución de posibles conflictos con los conocimientos que anteriormente se tenía.

3.2 CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) es una metodología abierta presentada en 1999 por las empresas NCR Systems Engineering Copenhagen (Estados Unidos y Dinamarca), DaimlerChrysler AG (Alemania), SPSS Inc. (Estados Unidos) y OHRA Verzekeringen en Bank Groep B.V. (Holanda). En la actualidad, esta metodología es una de las guías de referencia más utilizadas en proyectos de minería de datos (Moine *et al*., 2011). CRISP-DM establece un conjunto de tareas definidas en cuatros niveles de abstracción (fases, tareas generales, tareas específicas e instancias del proceso), que están estructuradas de forma jerárquica, iniciando desde el nivel general hasta el nivel específico (Chapman *et al*., 2000). El nivel superior está organizado por seis etapas (Figura 1): a) comprensión del negocio, b) comprensión de los datos, c) preparación de datos, d) modelado, e) evaluación, e f) implementación; donde cada una de estas etapas consta de tareas generales y específicas, esto con el objetivo de cubrir todas las acciones y decisiones relacionadas con el proyecto.

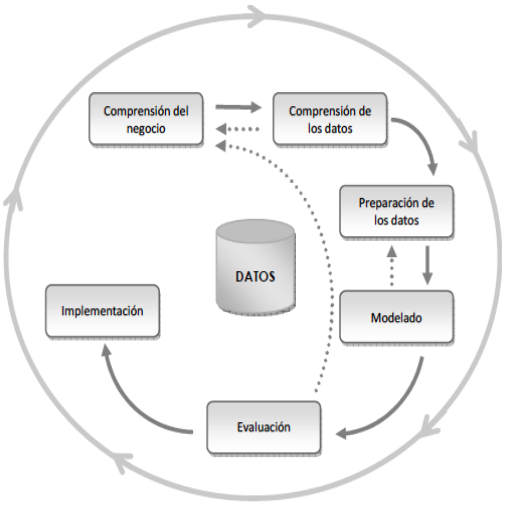

Figura 1. Etapas del proceso CRISP-DM. Fuente: Chapman et al. (2000).

1. Comprensión del negocio. Se centra en entender los objetivos y requerimientos del proyecto desde una perspectiva del negocio, con la finalidad de elaborar un plan preliminar para alcanzar los objetivos.

2. Comprensión de los datos. Consiste en la recolección y familiarización con los datos, para identificar problemas en la calidad de los mismos, por ejemplo, si existen datos repetidos, incompletos, inconsistentes, con errores, entre otros.

3. Preparación de los datos. Abarca todas las actividades para construir el conjunto de datos a utilizar. Las tareas de esta fase pueden ser realizadas en reiteradas ocasiones, ya sea a través de la limpieza de datos, la generación de variables adicionales, la integración de diferentes conjuntos de datos y cambios de formato.

4. Modelado. Se selecciona las técnicas apropiadas para la construcción de un prototipo. Al existir diversas técnicas con diferentes requisitos sobre los datos para un mismo problema, muchas veces es necesario volver a la etapa anterior para ajustar la vista de datos minable.

5. Evaluación. El modelo obtenido es evaluado con la finalidad de asegurarse que se logró alcanzar los objetivos iniciales del proyecto. Por lo que, esta fase concluye al aceptarse los resultados obtenidos.

6. Despliegue. En esta etapa el conocimiento adquirido es presentado al usuario final de manera tal que sea fácil de entender e interpretar, por ejemplo a través de un reporte.

CRISP-DM se caracteriza por hacer énfasis en los detalles de cada fase, es decir, cada etapa se divide en diferentes tareas y actividades. Por lo que, Rivo *et al*. (2011) indican que esta metodología hace que los proyectos, grandes o pequeños, de minería de datos sean rápidos de desarrollar, fiables y manejables.

3.3 SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) es una metodología creada por SAS Institute (Statistical Analysis Systems), quien la define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de interés (SAS, 1998). Este proceso consta de cinco etapas (Figura 2) necesarias para guiar el desarrollo de un proyecto de minería de datos (Moine, 2013). Sumathi y Sivanandam (2006) mencionan que la metodología SEMMA permite aplicar la estadística exploratoria y técnicas de visualización de manera fácil, así como la selección y transformación de las variables más significativa, con el objetivo de crear modelos para predecir resultados y evaluarlos de manera que sirva de apoyo para la toma de decisiones.

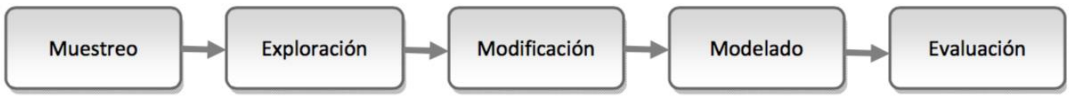

Figura 2. Etapas del proceso SEMMA. Fuente: Moine (2013).

1. Muestreo. En esta etapa se toma una muestra del conjunto de datos disponible, que debe ser lo suficientemente grande para contener la información relevante, y lo suficientemente pequeña como para correr el proceso rápidamente. Esta etapa es aconsejable cuando el tamaño del conjunto de datos es demasiado extenso.

2. Exploración. Consiste en explorar los datos en búsqueda de relaciones y tendencias desconocidas. Es una etapa especial para familiarizarse con los datos, y formular nuevas hipótesis a partir de su análisis.

3. Modificación. Etapa de preparación de datos que consiste en la limpieza de los valores atípico, se realiza un tratamiento de los datos faltantes, y se seleccionan, crean y modifican las variables que servirán para la etapa del modelado.

4. Modelado. Consiste en la creación del modelo para predecir las variables, utilizando algunas de las técnicas predictivas como árboles de decisión, redes neuronales, análisis discriminante o análisis de regresión.

5. Evaluación. En esta fase se evalúa la utilidad y la exactitud de los modelos obtenidos en el proceso de minería de datos, por ejemplo analizando la capacidad predictiva de los mismos.

Una clara diferencia con respecto a otras metodologías es que en SEMMA la primera fase se inicia con el muestreo de datos. Por otra parte, SEMMA está relacionada particularmente con productos comerciales de SAS Institute.

3.4 Catalyst

Es una metodología conocida como P3TQ (Product, Place, Price, Time, Quantity) conformada por dos modelos (Pyle, 2003): un modelo de negocio (MII) y un modelo de explotación de información (MIII). MII ofrece una guía para el desarrollo y construcción de un modelo con el objetivo de hacer frente a un problema u oportunidad de negocio. MIII proporciona una guía para la realización y ejecución de modelos de minería de datos con base en MII. Además, MII plantea cinco escenarios diferentes de acuerdo a las circunstancias del negocio (Britos, 2008):

a. Dato. El proyecto comienza con un conjunto de datos con el objetivo de explorarlos para encontrar patrones de interés.

b. Oportunidad. El proyecto inicia como un problema u oportunidad de negocio que debe ser explorada.

c. Prospectiva. El objetivo del proyecto es descubrir donde la minería de datos puede ofrecer un valor a la organización.

d. Definido. El proyecto comienza con la premisa de crear la especificación del modelo de minería de datos con un propósito específico.

e. Estratégico. El proyecto comienza con una estrategia de análisis para dar soporte a un escenario planificado por la organización.

Por su parte, MIII proporciona una guía de referencia para la explotación de información mediante una serie de pasos (Moine, 2013):

a. Preparación de los datos. Incluye una series de actividades que permiten comprobar la calidad de los datos a utilizar, se revisan las características de las variables, así como el tamaño de los datos, entre otros.

b. Selección de herramientas y modelado inicial. Permite seleccionar la herramienta y el modelo con base al análisis del problema, por ejemplo, si se van a predecir los datos es necesario conocer el tipo de tarea predictiva que más se ajuste.

c. Refinar el modelo seleccionado.

d. Implementar el modelo.

e. Comunicación de resultados. Presentar el resultado obtenido al público interesado y los responsables de la toma de decisiones.

3.5 Six Sigma

De acuerdo con Brady y Allen (2006) Six Sigma "*es un método organizado y sistemático para la mejora de procesos, nuevos productos y servicios basados en métodos estadísticos y científicos con el fin de reducir las tasas de defectos establecidos por el cliente*". Partiendo de esta definición, Six Sigma ha sido adoptado por diversas

empresas como un enfoque disciplinario para la resolución de problemas que involucra el análisis de datos, a través del empleo de herramientas estadísticas, con el fin de reducir la variación mediante la mejora continua (Jang y Jeon, 2009). Pyzdek y Keller (2003) señalan que Six Sigma es útil para el proceso de desarrollo de proyectos de minería de datos, puesto que está dirigido para mejorar la satisfacción del cliente, aumentar la calidad y reducir los costos y tiempos. Por tal motivo, Jang y Jeon (2009) proponen la integración de minería de datos en la metodología Six Sigma (Tabla 2).

Tabla 2. Actividades de minería de datos relacionadas con las etapas de Six Sigma.

Fases	Acción	Función	Beneficios
Medir	Manipulación de datos	Facilidad de manipular grandes volúmenes de datos (muestreo, partición y otros).	Manipulación eficiente de datos para mejorar la calidad de los datos.
Analizar	Análisis exploratorio	Análisis exploratorio utilizando métodos de visualización.	Fácil de explorar las diversas variables y un análisis gráfico interactivo.
	Descubrimiento de conocimiento	Descubrir conocimiento usando selección de variables, árboles de decisión, regresión, entre otros.	Una base solidad de la hipótesis estadística, de fácil interpretación, y la derivación de conocimiento utilizando un método fiable.
	Modelado	Uso de regresión, árboles de decisión y redes neuronales artificiales para elaborar modelos complejos.	Fácil para desarrollar un modelo más preciso y general.
		Selección del modelo que más se ajuste a diversas herramientas de evaluación.	Diversos modelo de evaluación y análisis gráfico.
Mejorar	Optimización	Encontrar el nivel límite de control de procesos a través del algoritmo IGN (nodo interactivo de agrupamiento).	Tener un control óptimo del rango establecido por el algoritmo de minería de datos.

Fuente: Jang y Jeon (2009).

La integración de actividades de minería de datos en la metodología Six Sigma se da con la finalidad de ofrecer al usuario un análisis de la calidad de datos en grandes cantidades de información.

4. Comparación de los procesos

Las metodologías presentadas comparten la misma filosofía, esto es, están estructuradas en diversas fases relacionadas entre sí con la finalidad de guiar el desarrollo de proyectos de minería de datos. Sin embargo, SEMMA parte del empleo del uso de la estadística (muestreo de datos), mientras que KDD, CRISP-DM, Catalyst y Six Sigma se centran en el análisis de los requerimientos y el entendimiento del negocio. A su vez, cada una de estas metodologías contemplan tareas específicas para el entendimiento, selección y preparación de datos; así como la aplicación de los algoritmos para el descubrimiento de patrones de interés. Asimismo, la etapa de evaluación forma parte crucial en todas las metodologías, esto debido a la importancia de validar los resultados obtenidos, por ejemplo, SEMMA y Six Sigma interpretan y evalúan los resultados con base en el desempeño de modelo, mientras que en KDD y Catalyst la validación está en función de los objetivos del proyecto. Para el caso de CRISP-DM los resultados se evalúan con base en el desempeño del modelo y el cumplimiento de los requerimientos iniciales del proyecto. Otro aspecto a considerar sobre estas metodologías es el uso de herramientas empleadas para el desarrollo de proyectos de minería de datos, por ejemplo, SEMMA está relacionada con productos comerciales de SAS Institute, como Enterprise Miner y Text Miner. Esto trae como consecuencia que el analista de datos tenga que ajustarse a los algoritmos y herramientas de la misma. Además, dado que KDD, CRISP-DM y Catalyst fueron diseñadas como metodologías neutras, de libre distribución, es decir, sin costo, éstas pueden adaptarse a cualquier herramienta ya sea libre o comercial. En la Tabla 3 se presenta un cuadro comparativo con las principales características de las cinco metodologías presentadas. Se incluye sus etapas, el tipo de herramientas utilizadas, el objetivo de su evaluación, el año de su creación, entre otros aspectos.

Tabla 3. Comparativa entre las metodologías tradicionales de minería de datos.

	KDD	CRISP-DM	SEMMA	Catalyst	Six Sigma
Fases	• Integración y recopilación • Selección, limpieza y transformación • Minería de datos • Evaluación e interpretación • Difusión y uso	• Entendimiento del negocio • Entendimiento de los datos • Preparación de los datos • Modelado • Evaluación • Despliegue	• Muestreo • Exploración • Modificación • Modelado • Evaluación	• Preparación de los datos • Modelado • Refinar el modelo • Implementar el modelo • Comunicación de resultados	• Definición • Medición • Análisis • Mejora • Control
Etapas iterativas	Si	Si	No	Si	No
Elección de herramientas	Libres y comerciales	Libres y comerciales	Comerciales	Libres y comerciales	Libres y comerciales
Tipo de evaluación del resultado	Basado en los objetivos del proyecto	Basado en el modelo y los objetivos del proyecto	Basado en el modelo	Basado en los objetivos del proyecto	Basado en el modelo
Diseñada para minería de datos	Si	Si	Si	Si	No
Año de publicación	1996	1999	1998	2003	1986

Fuente: Creación propia.

Con base en la Tabla 3, los procesos de minería de datos han evolucionado con el paso de los años, esto con el objetivo de cumplir con los requerimientos definidos en los proyectos. Tal es así que en la década de los 80 nace Six Sigma orientado al análisis de datos (1986) cuyo propósito es reducir la variación mediante la mejora continua de los procesos. Posteriormente, en la década de los 90 surgieron los procesos KDD (1996), SEMMA (1998) y CRISP-DM (1999), los cuales comparten características similares para la explotación de información; siendo CRISP-DM una de las

metodologías más utilizadas debido al nivel de detalle que presenta cada una de sus etapas. Siguiendo esta filosofía, en el 2003 surge Catalyst que permite describir a detalle cada una las etapas del proceso de minería de datos, haciendo énfasis en aspectos organizacionales, oportunidades de negocio y a la necesidad de incluir al usuario. Sin embargo, a pesar de que estas metodologías cumplen con el objetivo principal de guiar el descubrimiento de patrones de intereses en volúmenes de datos, aun carecen de aspectos importantes como es la integración del usuario como elemento principal y la visualización eficiente de los patrones obtenidos. Ambos aspectos son cruciales para una mejor explicación y entendimiento en la generación del nuevo conocimiento. Por lo que, en este trabajo de investigación se hace una propuesta, como marco de referencia, para el desarrollo de proyectos de minería de datos centrada en el usuario.

5. Propuesta de minería de datos centrado en el usuario

Los procesos analizados cumplen con el propósito de encaminar el desarrollo de un proyecto de minería de datos; sin embargo, éstas no consideran al usuario como factor principal en sus etapas. Por lo que, surge la necesidad de disponer de un proceso de minería de datos centrado en el usuario, que sirva como marco de referencia para el desarrollo de proyectos de explotación de datos con el fin de alcanzar los objetivos de manera satisfactoria. En este sentido, para este nuevo marco metodológico se recurre al diseño centrado en el usuario, a través de la norma ISO 9241-210, y el proceso CRISP-DM como guía de referencia para llevar a cabo proyectos de minería de datos. En la actualidad, ambos procesos son las principales guías de referencia a nivel internacional. Por tanto, a través de integración de ambos procedimientos se pretende proporcionar al usuario una nueva e innovadora metodología que mejore el proceso de análisis y explotación de datos para el reconocimiento de patrones.

5.1 ISO 9241-210:2010 (Human-centered design for interactive systems)

Los últimos años el diseño centrado en el usuario ha tenido una notable aceptación debido al uso procesos que encaminan el diseño de aplicaciones que respondan a las necesidades reales de los usuarios finales (Sánchez, 2011). Por tal motivo, se han definido modelos y estándares como guía de referencia para el diseño centrado en el usuario. Uno de estos estándares es la norma ISO 9241-210:2010, que fue definida por la Organización Internacional de Normalización (ISO), específicamente por el comité de ergonomía de la interacción humano-computadora. Esta norma proporciona requisitos y recomendaciones para los principios de diseño centrado en usuario y actividades durante todo el ciclo de vida de los sistemas interactivos. Las etapas que comprende la norma ISO 9241-210:2010 son iterativos que involucran al usuario en todo el ciclo de desarrollo, estas son (Figura 3): a) el análisis del contexto de uso, b) la especificación de requerimientos, c) el diseño, d) la evaluación del diseño, y e) la solución de diseño.

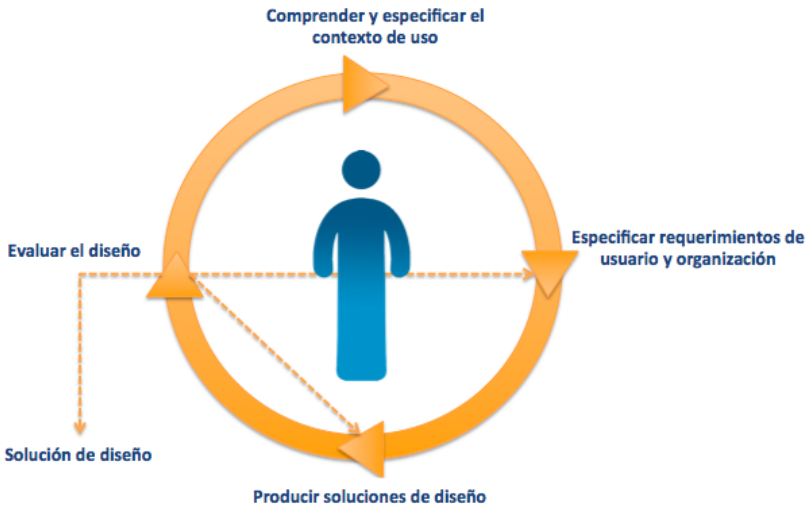

Figura 3. Iteración entre etapas de la norma ISO 9241-210. Fuente: Adaptado de ISO 9241-210:2010

En la Figura 4 se presenta el vínculo de las etapas que conforman la norma ISO 9241-210 y las fases del proceso CRISP-DM. A través de esta matriz se identifican las etapas significativas que se incluyen en la nueva metodología de minería de datos centrada en el usuario. Por ejemplo, la etapa de *especificación del contexto de uso* de la norma ISO 9241-210 se asocia al *entendimiento del negocio* de CRISP-DM. Esto representa el entendimiento de los objetivos del proyecto o negocio, tomando el cuenta las necesidades de los usuarios involucrados. Asimismo, la etapa de *especificación de requerimientos de usuario* se asocia al *entendimiento y preparación de datos*, que es la fuente de información utilizada para la extracción de patrones de interés; se define también las funciones que realizará el usuario (analista de datos) en el sistema.

ISO 9241-210 / *CRISP-DM*	Especificación del contexto de uso	Especificación de requerimientos de usuario	Producción de soluciones de diseño	Evaluación del diseño	Solución de diseño
Entendimiento del negocio					
Entendimiento de los datos					
Preparación de los datos					
Modelado					
Evaluación					
Implementación					

Fuente: Elaboración propia.

Figura 4. Asociación de las etapas de la norma ISO 9241-210:2010 y CRISP-DM.

Siguiendo lo anterior, la etapa de *producción de soluciones de diseño* se relaciona con la etapa de *modelado* de CRISP-DM. Caso similar ocurre con la *evaluación del diseño* que se asocia con la *evaluación del modelo*, cuyo propósito es validar el cumplimiento de los objetivos iniciales del proyecto y los resultados obtenidos. Otra de las etapas representativas que también se relacionan son el *despliegue* (CRISP-DM) y la *solución del diseño* (ISO 9241-210:2010), siendo útil para la visualización y representación de los patrones extraídos, así como también para la presentación de los resultados a los

usuarios interesados, como es el caso de los tomadores de decisiones.

Tomando como base la relación existente, en la Figura 5 se presenta el diseño conceptual de la nueva metodología para el desarrollo de proyectos de minería de datos centrada en el usuario. El objetivo de esta nueva metodología es involucrar al usuario en etapas significativas del proceso de descubrimiento de conocimiento, siguiendo para esto un ciclo iterativo, dividida en tres etapas generales (análisis, diseño y despliegue), que a su vez contienen etapas secundarias o sub-etapas, como: a) análisis contextual, análisis de datos y preparación de datos; b) prototipado, modelado y evaluación; y c) visualización; respectivamente.

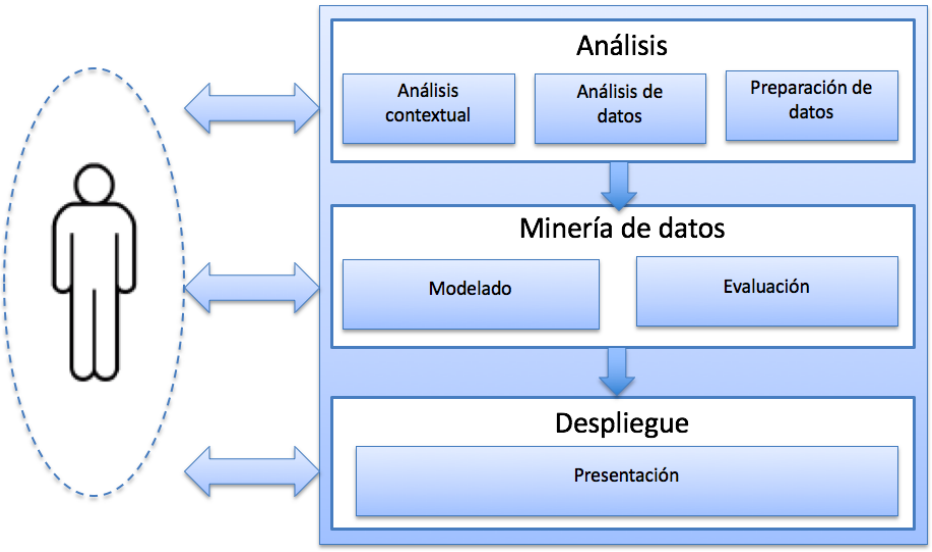

Figura 5. Propuesta de un proceso de minería de datos centrado en el usuario.

En esta propuesta se da prioridad e importancia al usuario, por lo que es necesario conocer sus gustos, objetivos, necesidades, actividades, entorno de trabajo, entre otros aspectos. Esta integración del usuario como elemento principal hace que este marco de referencia tenga ventajas sobre otras metodologías tradicionales existentes.

6. Conclusiones

A pesar de la amplia variedad de tareas y técnicas de minería de datos, es necesario definir un marco de trabajo que permita planificar y guiar el proceso de desarrollo de un proyecto de minería de datos centrado en el usuario.

Existen procesos tradicionales de minería de datos que han evolucionado con el paso de los años, pero estos no incluyen al usuario como factor significativo que debe considerarse para el éxito de proyectos de explotación de datos. Además, estos procesos carecen de una visualización eficiente de los patrones obtenidos.

Para esta propuesta de minería de datos centrado en el usuario se toma como base dos proceso internacionales ampliamente conocidos CRISP-DM como guía de referencia para el proceso de reconocimiento de patrones y la norma ISO 9241-210:2010 para la construcción de sistemas interactivos centrados en el usuario.

Finalmente, como trabajo futuro se tiene poner en práctica el marco metodológico de minería de datos centrado en el usuario aplicado a un caso de estudio para el análisis y explotación de fuentes de datos, mejorando la interacción del usuario en el proceso de descubrimiento de conocimiento y la visualización de los patrones.

Referencias

Brady J. E. y Allen T. T. (2006). Six Sigma Literature: A Review and Agenda for Future Research. *Quality and Reliability Engineering International*, 22(3), 335-367.

Britos P. V. (2008). Procesos de explotación de información basados en sistemas inteligentes (Tesis doctoral). *Universidad Nacional de la Plata*, Buenos Aires, Argentina.

Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C. y Wirth R. (2000). CRISP-DM 1.0 Step-by-step Data Mining Guide. <www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf>. Última consulta 08.07.2015.

Fayyad U., Piatetsky-Shapiro G., y Smyth P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-54.

Govindarajan M. y Chandrasekaran R. M. (2011). Intrusion detection using neural based hybrid classification methods. *Computer networks*, 55(8), 1662-1671.

Hernández J., Ramírez M. J. y Ferri C. (2004). Introducción a la Minería de Datos. Pearson Educación. Editorial Pearson Prentice Hall, pp. 680, ISBN: 84-205-4091-9, Madrid, España.

ISO 9241-210 (2010). Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO)*.

Jang G. S. y Jeon J. H. (2009). A Six Sigma Methodology Using Data Mining: A Case Study on Six Sigma Project for Heat Efficiency Improvement of a Hot Stove System in a Korean Steel Manufacturing Company. *Cutting-Edge Research Topics on Multiple Criteria Decision Making*, 72-80.

Larose D. T. (2014). Discovering knowledge in data: an introduction to data mining. John Wiley y Sons, pp. 336, New Jersey.

MIT (2001). The Technology Review Ten, MIT Technology Review, January/February. <www.techreview.com>. Última consulta 25.08.2015.

Moine J. M., Gordillo S. y Haedo A. (2011). Análisis comparativo de metodologías para la gestión de proyectos de Minería de Datos. VIII Workshop Bases de Datos y Minería de Datos. 931-938.

Moine J. M. (2013). Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo (Tesis doctoral). Universidad Nacional de la Plata, Buenos Aires, Argentina.

Molero G. y Céspedes Y. (2014). Data Mining and Knowledge Discovery: An Introduction. Capítulo de libro Knowledge Discovery in Databases. Ed. Academy Publish (en prensa).

Pyle D. (2003). Business modeling and data mining. Ed. Morgan Kaufmann, pp. 720, ISBN: 978-1558606531.

Pyzdek T. y Keller P. A. (2003). The six sigma handbook. Editorial McGraw-Hill, pp. 848, ASIN: B000SEGKDY, New York.

Rivo E., de la Fuente J., Rivo Á., García E., Cañizares M. y Gil P. (2012). Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology*, 14(1), 73-79.

Sánchez W. O. (2011). La usabilidad en Ingeniería de Software: definición y características. *Ing-novación*. 2, 7-22.

SAS Institute (1998). Data Mining and the Case for Sampling. Data Mining Using SAS Enterprise Miner. <http://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf>. Última consulta 20.06.2015.

Sumathi S. y Sivanandam S. (2006). Introduction to Data Mining and its Applications. *Studies in Computational Intelligence*, 29, editado por Springer-Verlag, pp. 828, ISBN: 3-540-34350-4, Heidelberg, Alemania.

Tan P. N., Steinbach M., y Kumar V. (2006). Introduction to data mining. Boston: Pearson Addison Wesley, pp. 769, ISBN: 978-0321321367.