Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM

EPB 603 Sistemas del Conocimiento

Basado en la Tesis: "Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM)" de José Alberto Gallardo Arancibia.

Modelos de proceso para proyectos de Data Mining (DM)

Son diversos los modelos de proceso que han sido propuestos para el desarrollo de proyectos de Data Mining tales como SEMMA *(Sample, Explore, Modify, Model, Assess)* [SAS, 2003], DMAMC *(Definir, Medir, Analizar, Mejorar, Controlar)* [Isixsigma, 2005], o CRISP-DM *(Cross Industry Standard Process for Data Mining)* [CRISP-DM, 2000], sin embargo uno de los modelos principalmente utilizados en los ambientes académico e industrial es el modelo CRISP-DM.

CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP–DM [CRISP-DM, 2000], es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, como se puede constatar en la gráfica presentada en la figura 2.3. Esta gráfica, publicada el año 2007 por kdnuggets.com, representa el resultado obtenido en sucesivas encuestas efectuadas durante los últimos años, respecto del grado de utilización de las principales guías de desarrollo de proyectos de Data Mining. En ella se puede observar, que a pesar de que el uso de aun frente a otras, la guía de referencia más ampliamente utilizada.

Los orígenes de CRISP-DM, se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases) [Reinartz, 1995], [Adraans, 1996], [Brachman, 1996], [Fayyad, 1996], el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining).

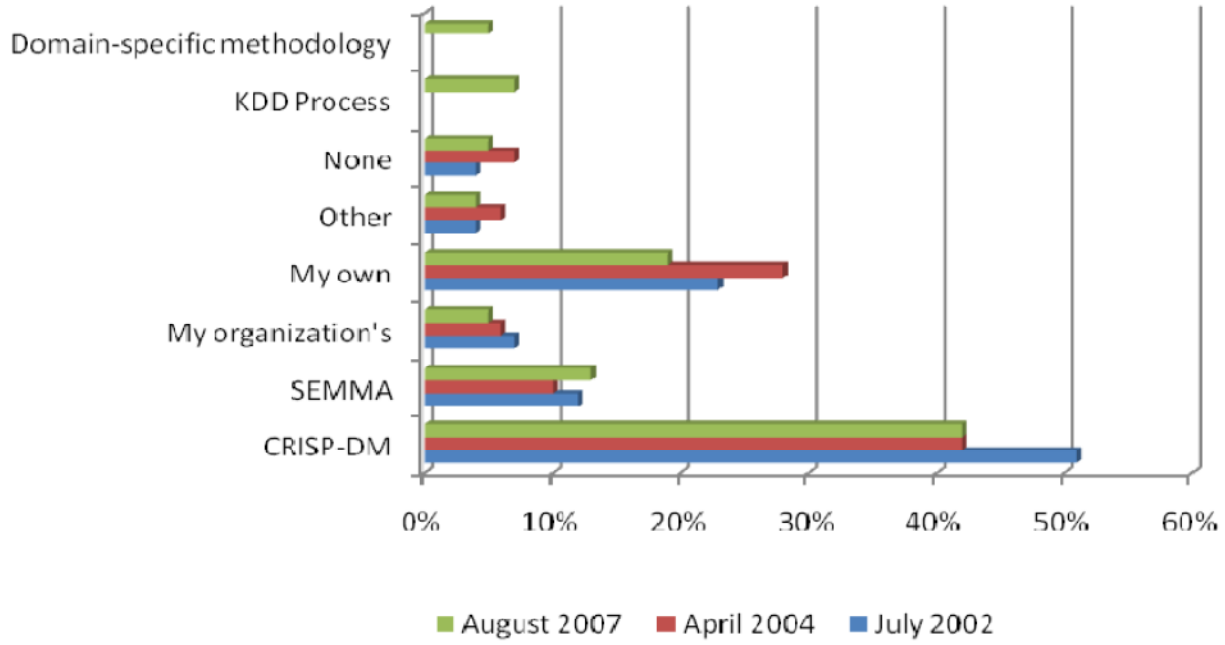

Figura No. 2.3. Metodologías utilizadas en Data Mining ([kdnuggets, 2007]).

CRISP-DM, está dividida en 4 niveles de abstracción organizados de forma jerárquica (figura 2.4) en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Data Mining, en una serie de seis fases (figura 2.5):

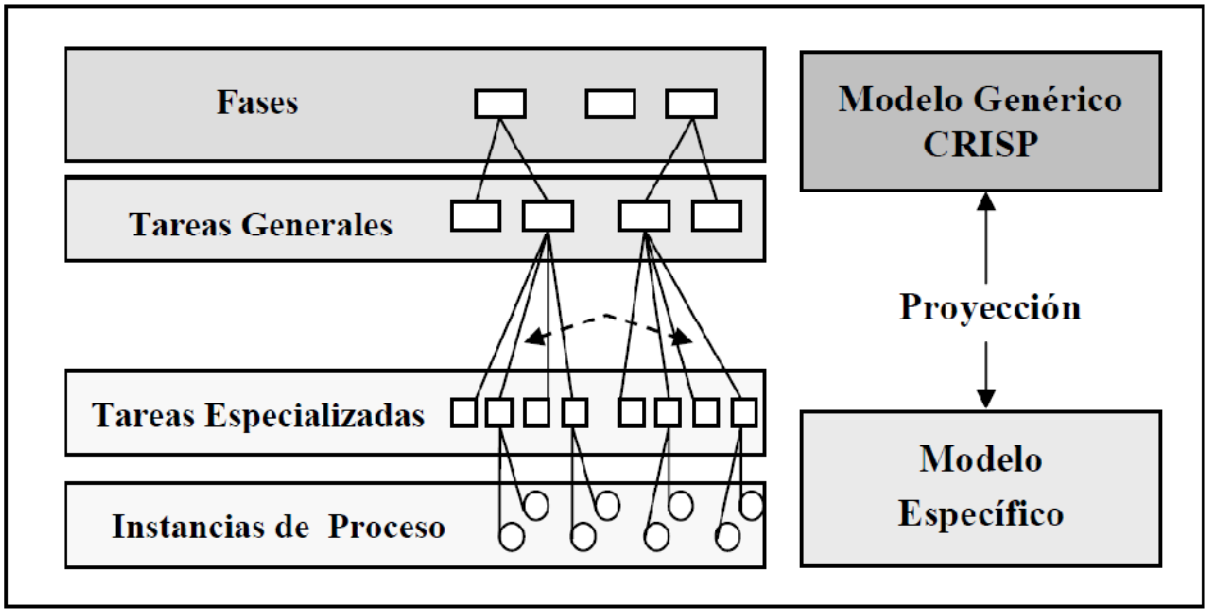

Figura No. 2.4. Esquema de los 4 niveles de CRISP-DM ([CRISP-DM, 2000]).

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

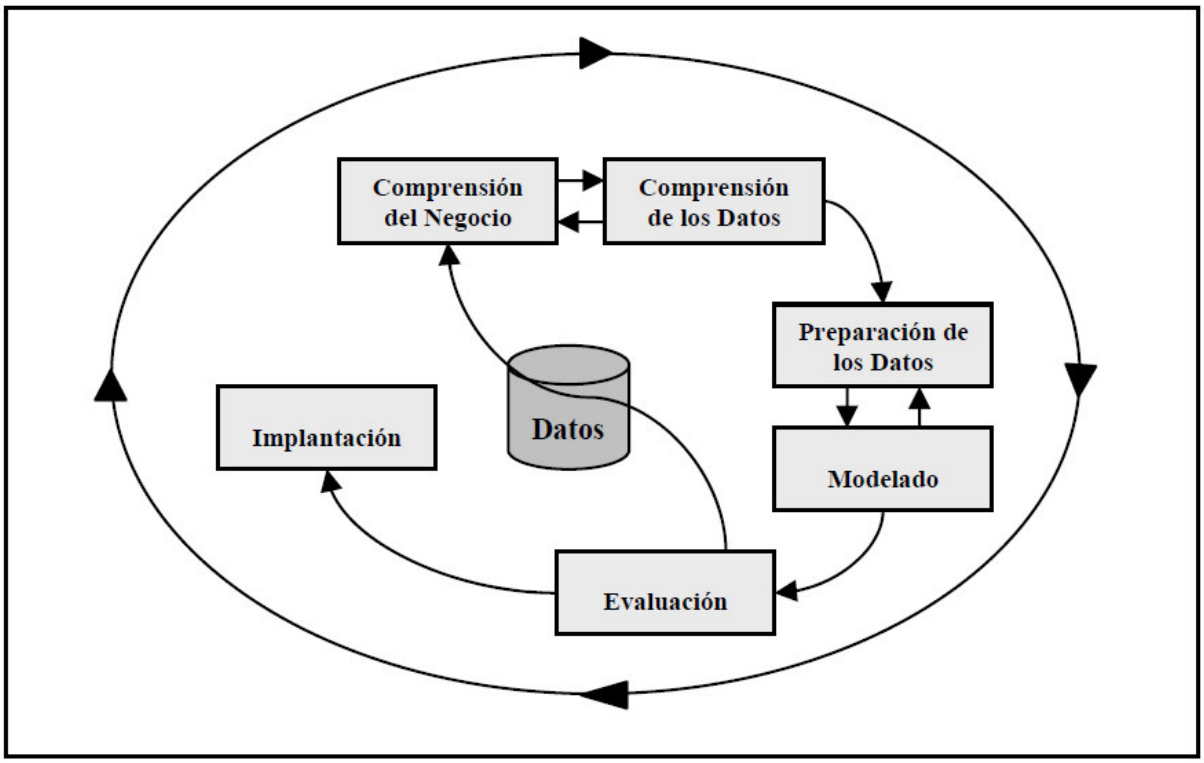

Figura No. 2.5. Modelo de proceso CRISP–DM ([CRISP-DM, 2000]).

A continuación se describen cada una de las fases en que se divide CRISP-DM.

1. Fase de comprensión del negocio o problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema (figura 2.6), es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

Determinar los objetivos del negocio. Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito. Los problemas pueden ser diversos como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio,

califica el resultado del proceso de DM, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

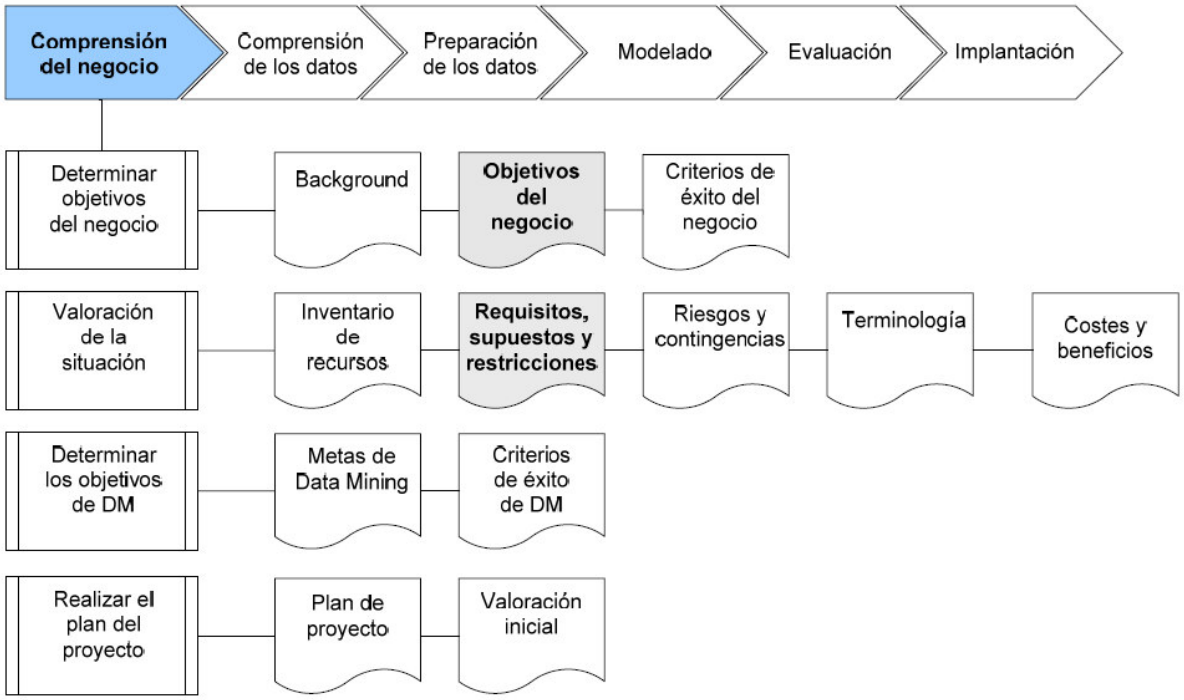

Figura No. 2.6. Fase de comprensión del negocio ([CRISP-DM, 2000]).

Evaluación de la situación. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de DM?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

Determinación de los objetivos de DM. Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento. *Producción de un plan del proyecto.* Finalmente esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

2. Fase de comprensión de los datos

La segunda fase (figura 2.7), fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las

primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

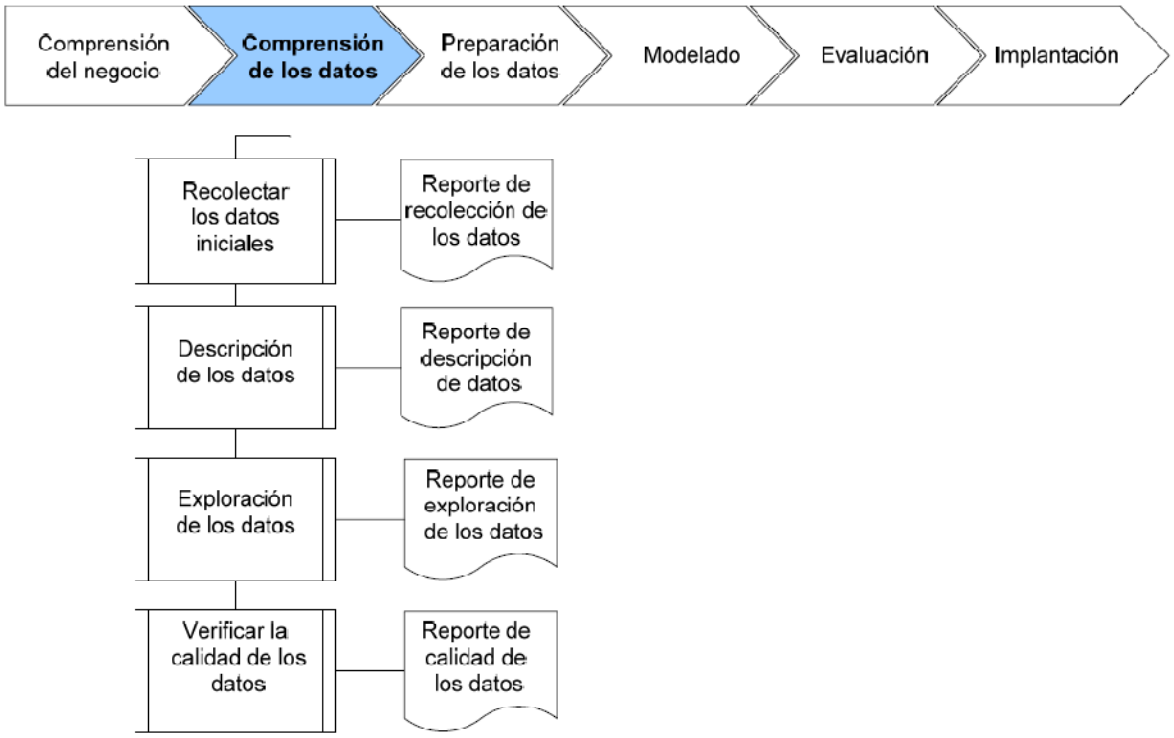

Figura No. 2.7. Fase de comprensión de los datos ([CRISP-DM, 2000]).

Las principales tareas a desarrollar en esta fase del proceso son:

Recolección de datos iniciales. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

Descripción de los datos. Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

Exploración de datos. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

Verificación de la calidad de los datos. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos.

3. Fase de preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. La figura 2.8, ilustra las áreas de que se compone ésta, e identifica sus salidas. Una descripción de las tareas involucradas en esta fase es la siguiente: *Selección de datos.* En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas. *Limpieza de los datos.* Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.

Estructuración de los datos. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

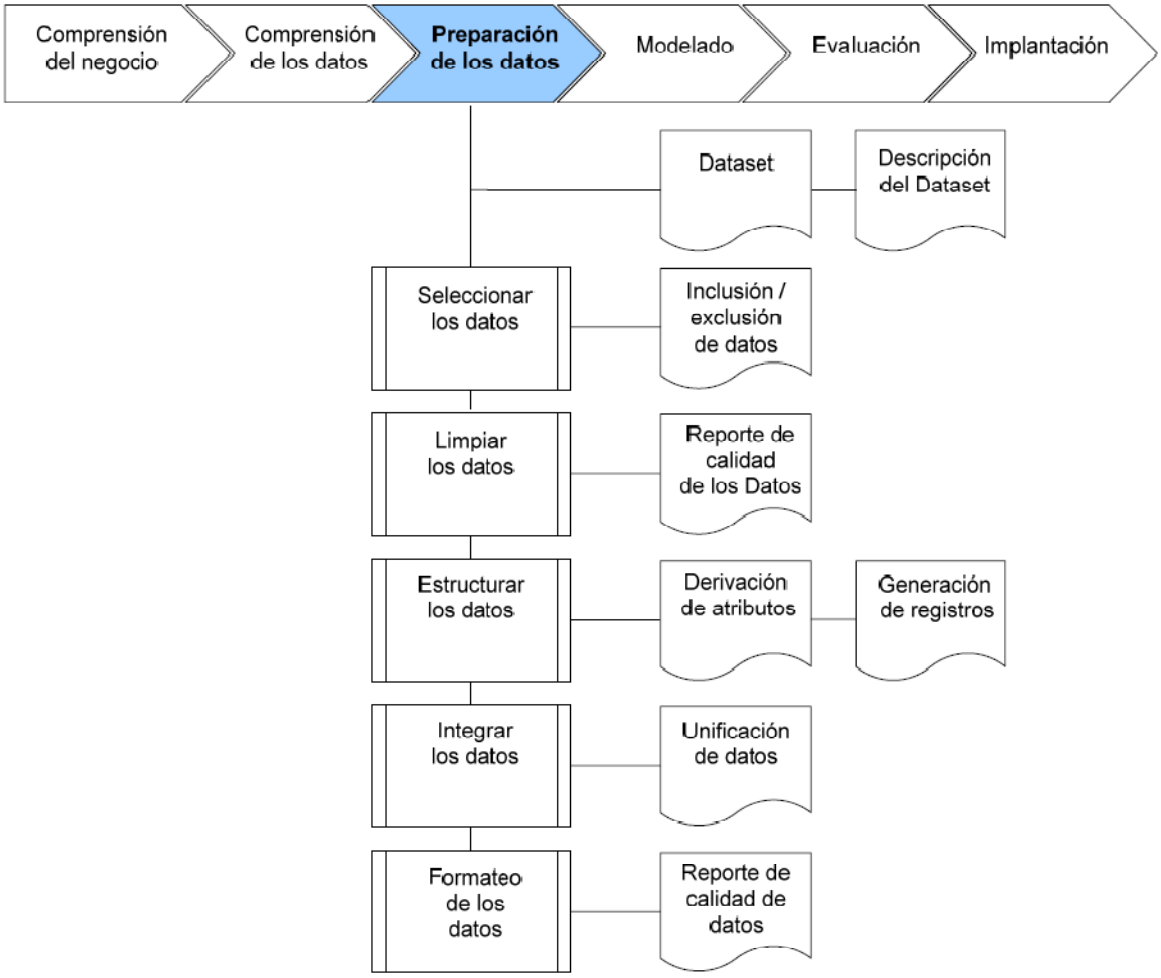

Figura No. 2.8. Fase de preparación de los datos ([CRISP-DM, 2000]).

Integración de los datos. La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

Formateo de los datos. Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

4. Fase de modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

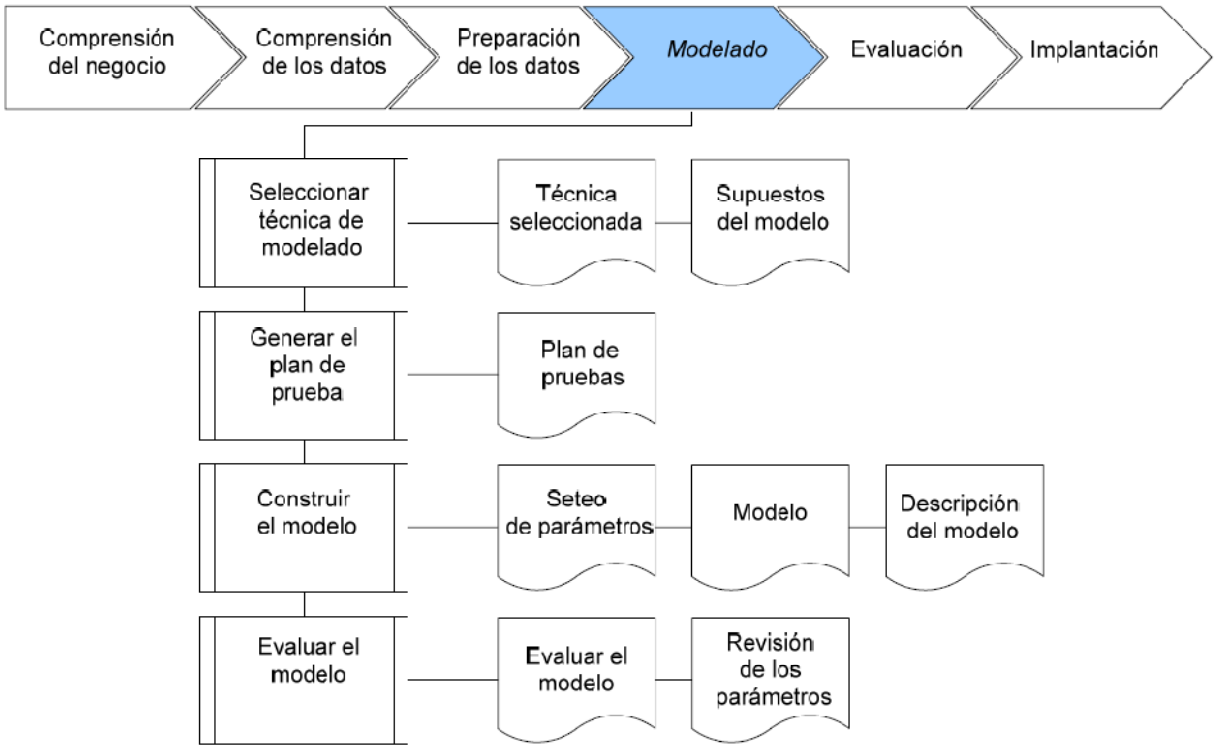

Figura No. 2.9. Fase de modelado ([CRISP-DM, 2000]).

Previamente al modelado de los datos, se debe determinar un método de evaluación de los odelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo. La figura 2.9 ilustra las tareas y resultados que se obtienen en esta fase. Una descripción de las principales tareas de esta fase es la siguiente:

Selección de la técnica de modelado. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes. Por ejemplo, si el

problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

Generación del plan de prueba. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

Construcción del Modelo. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

Evaluación del modelo. En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc...).

5. Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Las *matrices de confusión* [Edelstein, 1999] son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. La figura 2.10 detalla las tareas que componen esta fase y los resultados que se deben obtener. Las tareas involucradas en esta fase del proceso son las siguientes:

Evaluación de los resultados. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

Proceso de revisión. El proceso de revisión, se refiere a calificar al proceso entero de DM, a objeto de identificar elementos que pudieran ser mejorados.

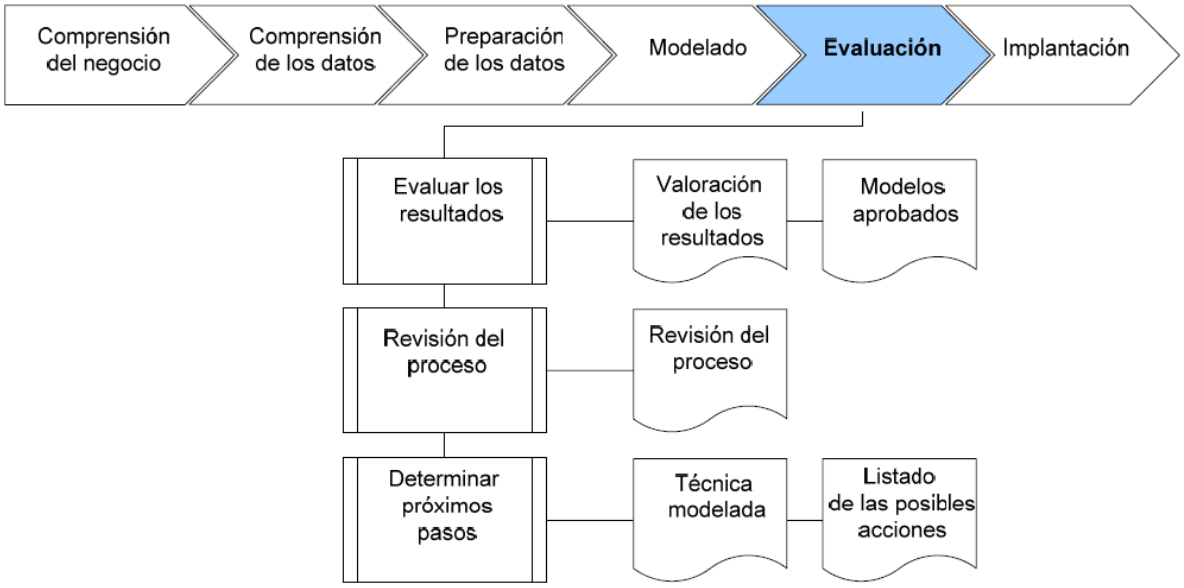

Figura No. 2.10. Fase de evaluación ([CRISP-DM, 2000]).

Determinación de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de DM.

6. Fase de implementación

En esta fase (figura 2.11), y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que se ejecutan en esta fase son las siguientes:

Plan de implementación. Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación. *Monitorización y Mantenimiento.* Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema

como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

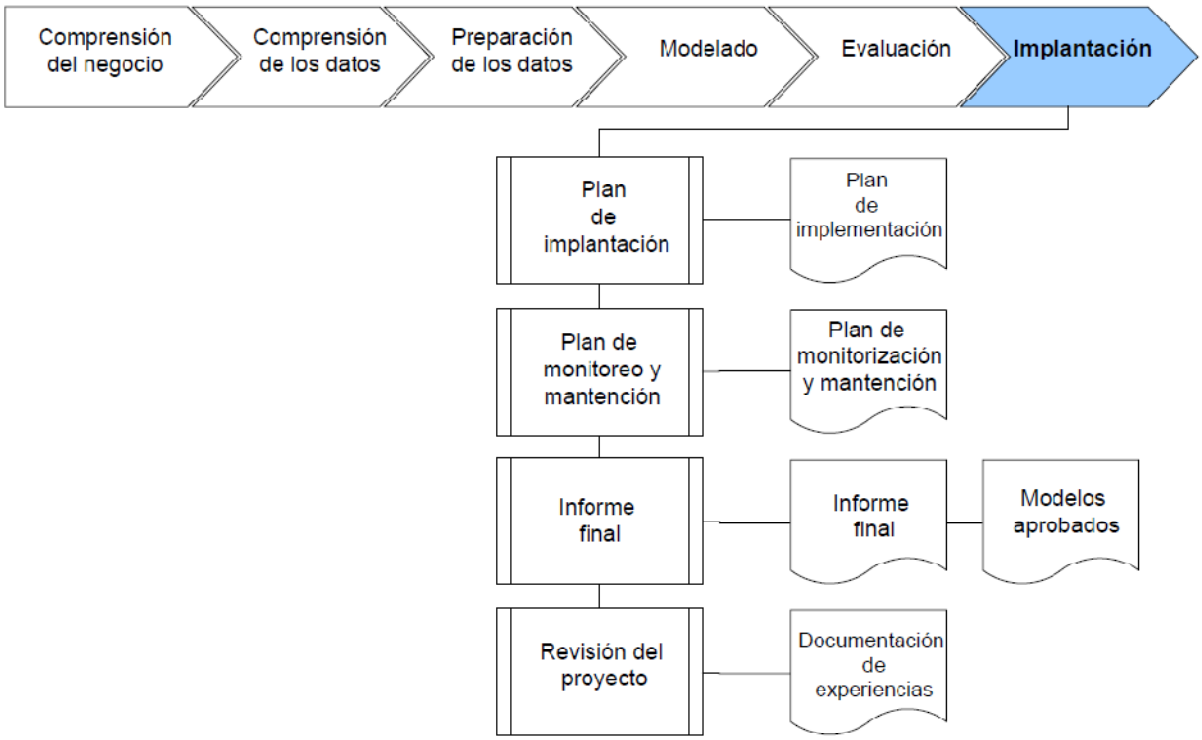

Figura No. 2.11. Fase de implementación ([CRISP-DM, 2000]).

Informe Final. Es la conclusión del proyecto de DM realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto. *Revisión del proyecto:* En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.