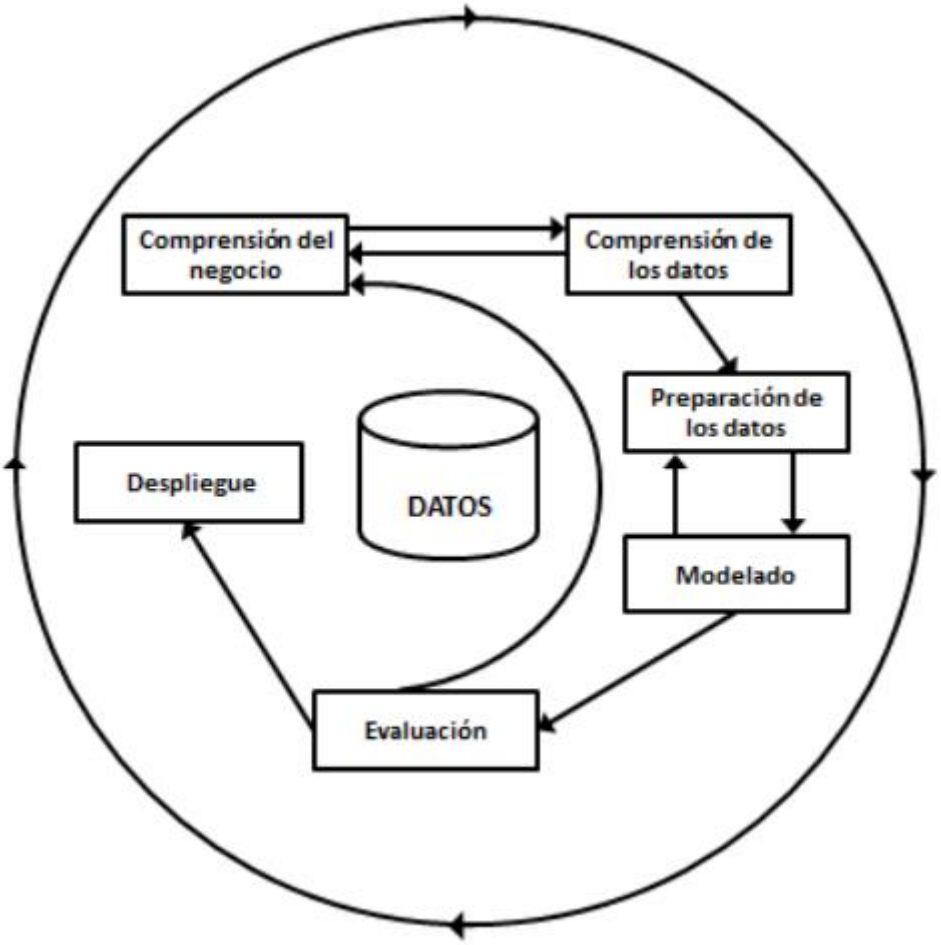
**Metodología CRISP-DM**

CRISP-DM (Cross Industry Standard Process for Data Mining), es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema.

Para implementar una tecnología en un negocio es necesaria una metodología. Estos métodos suelen venir de las experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos de implementación de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM.

CRISP-DM incluye un modelo y una guía, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene por qué ser ordenada desde la primera hasta la última. En la siguiente figura se puede observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

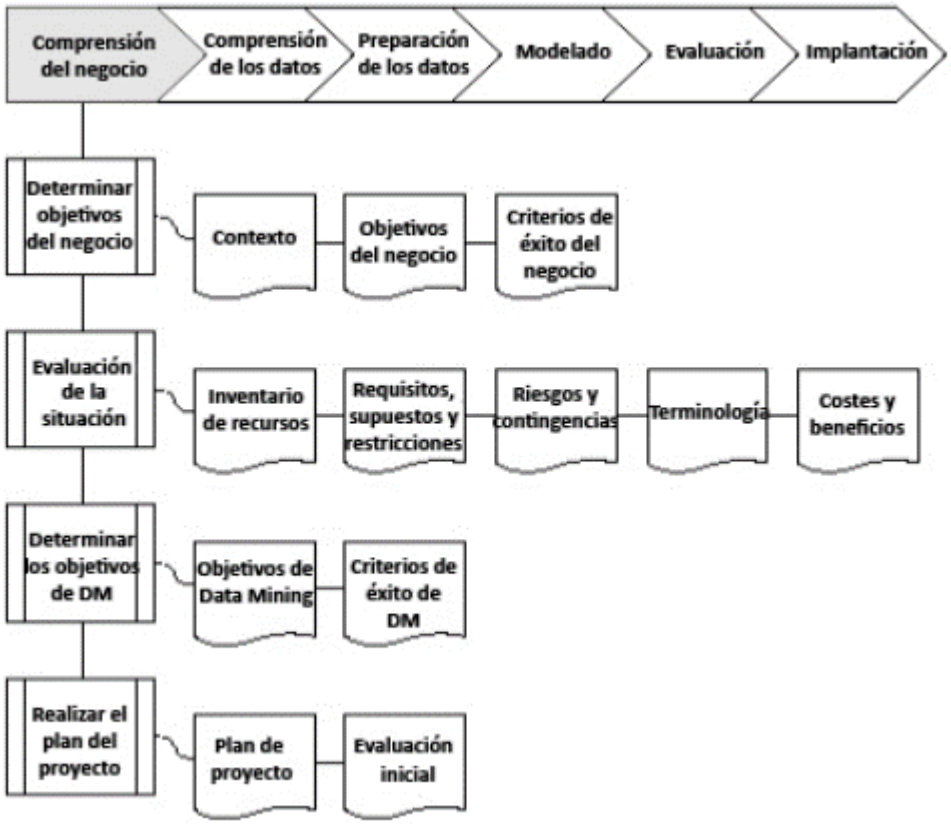



A continuación, se explica cada una de estas fases:

1. **Comprensión del negocio.**

Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, 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 la minería de datos, 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 minería de datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. A continuación vemos una descripción de cada una de las principales tareas que componen esta fase:




- **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 la minería de datos 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 minería de datos, o bien de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

- **Evaluación de la situación.**

  En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de minería de datos, 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 minería de datos?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de minería de datos.

- **Determinar los objetivos de la minería de datos.**

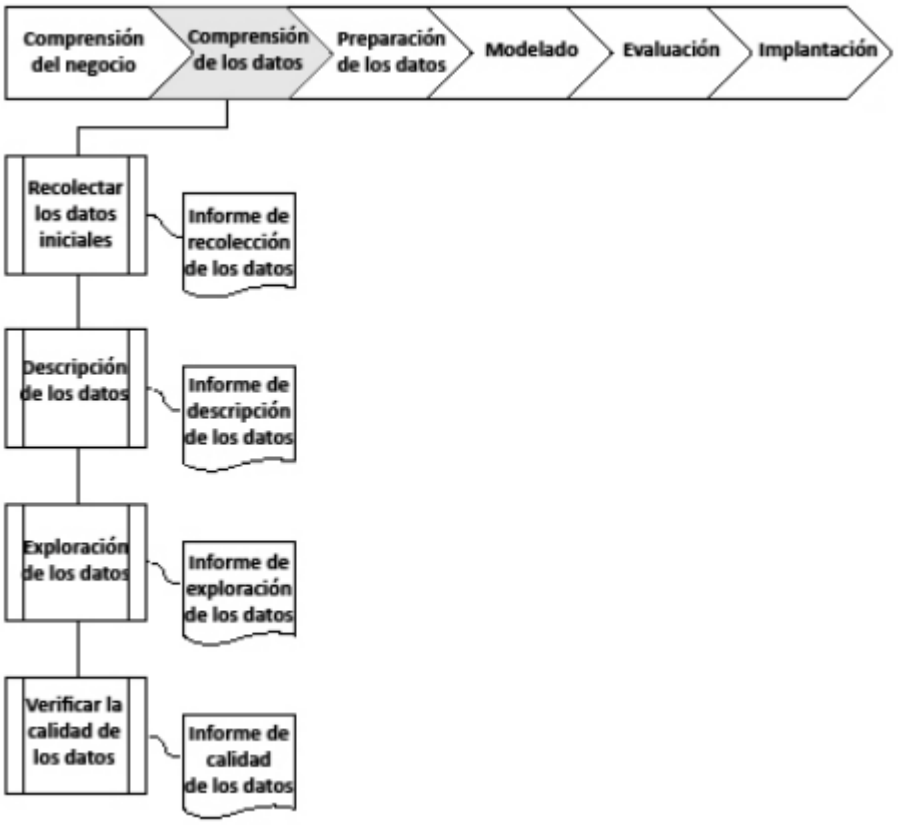
  Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de minería de datos, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de minería de datos será por ejemplo determinar el perfil de los clientes respecto de su capacidad de endeudamiento.

- **Realizar el plan del proyecto.**

  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 uno de ellos.

2. **Comprensión de los datos.**

   Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (Data Mining), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas. Vemos las tareas que componen esta fase.




- **Recolectar los 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 implica 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 los datos.**

  Una vez realizada la descripción de los datos, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto implica 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.

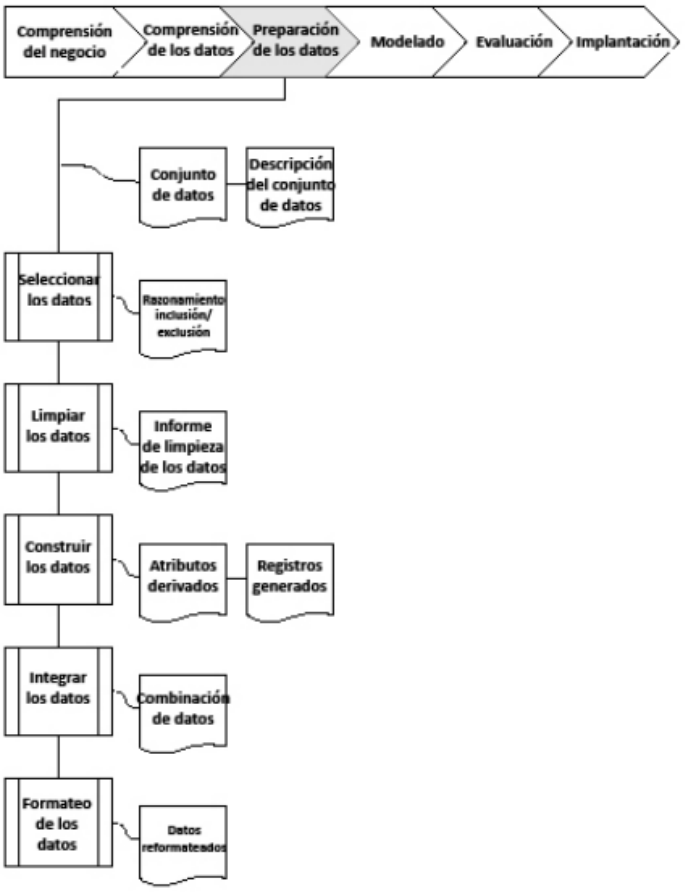
- **Verificar la calidad de los datos.**

  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 una vez llegados a este punto es poder garantizar la completitud y corrección de los datos.

3. **Preparación de los datos.**

   En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los 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, ya que en función de la técnica de modelado elegida, los datos requieren ser procesados de una manera o de otra, por esta razón las fases de preparación y de modelado interactúan de forma permanente. En la figura 7 [CRISP-DM, 2000] se pueden ver cada una de las tareas de las que se compone esta fase, así como las salidas de cada una de ellas.




- **Seleccionar los datos.**

  En esta etapa se selecciona un subconjunto de los datos adquiridos anteriormente apoyándose en criterios previamente definidos en las fases anteriores como la calidad de los datos en cuanto a su completitud, corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionados con las técnicas de minería de datos seleccionadas.

- **Limpiar 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 la normalización de los datos, discretización de campos numéricos, tratamiento de valores faltantes, reducción del volumen de datos, etc.

- **Construir los datos.**

  Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario - 28 - 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.

- **Integrar los datos.**

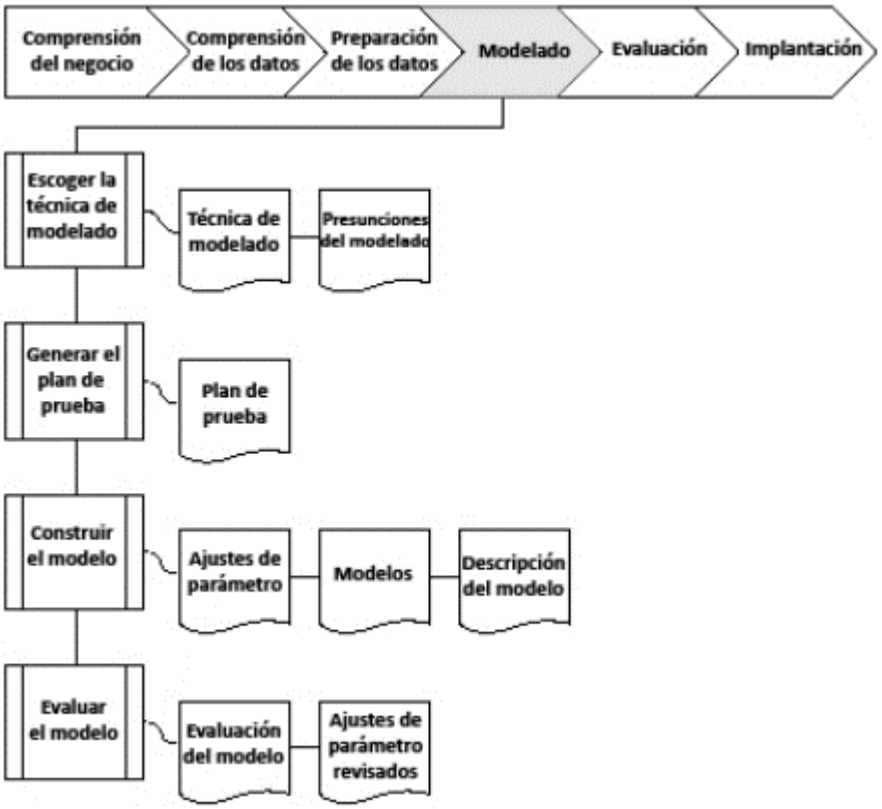
  La integración de los datos implica la creación de nuevas estructuras a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros 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 de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos en concreto, como por ejemplo la reordenación de los campos y/o de los 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. **Modelado.**

   En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios: o Ser apropiada para el problema. o Disponer de los datos adecuados. o Cumplir los requisitos del problema. o Tiempo adecuado para obtener un modelo. o Conocimiento de la técnica. Previamente al modelado de los datos se debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno 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 8 muestra las tareas y las salidas que se obtienen en esta fase, a continuación, describimos las tareas principales de esta fase.




- **Escoger la técnica de modelado.**

  Esta tarea consiste en la selección de la técnica de minería de datos más apropiada al tipo de problema que se quiere resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbors o razonamiento basado en casos (CBR), si el problema es de predicción, análisis de regresión o redes neuronales, o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

- **Generar el plan de prueba.**

  Se debe generar un procedimiento destinado a probar la calidad y validez del modelo elegido una vez que éste esté construido. Por ejemplo, en una tarea supervisada de minería de datos 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 Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario - 30 - luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

- **Construir el modelo.**

  A continuación, se ejecuta la técnica seleccionada 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

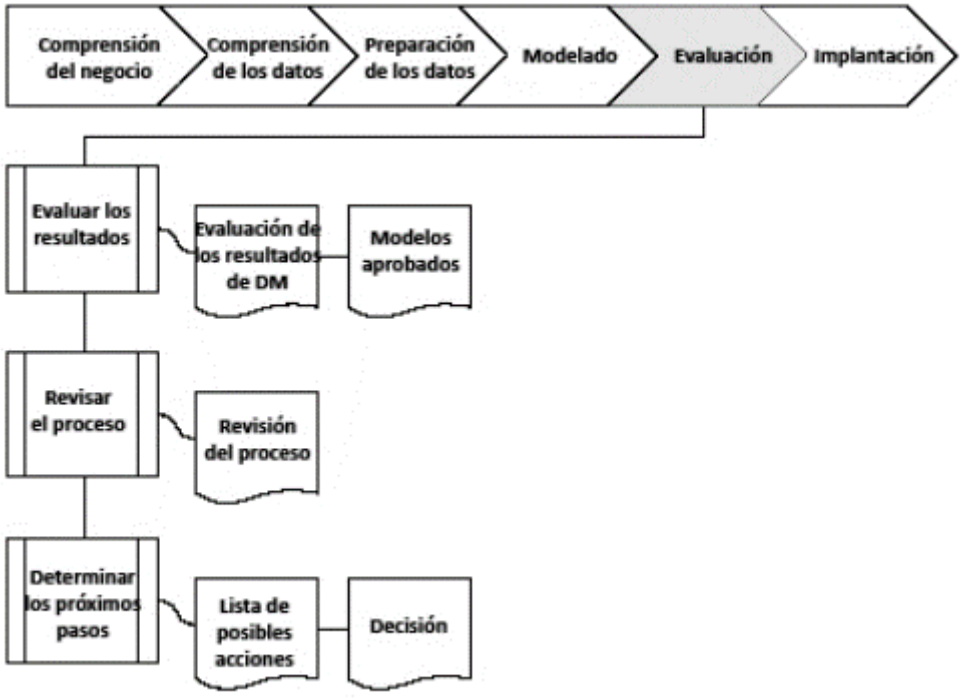
es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

- **Evaluar el modelo.**

  En esta última tarea de esta fase de modelado 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 minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.).

**5. 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 pueda haber cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. 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 9 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.




- **Evaluar los resultados.**

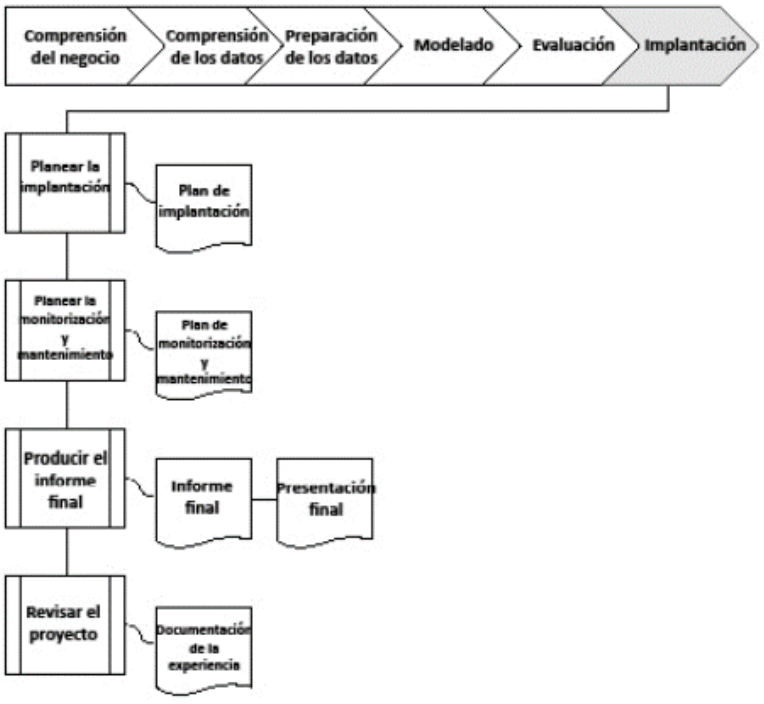
  En los pasos de evaluación anteriores se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual el modelo sea deficiente, o si es aconsejable probar el modelo en un problema real si el tiempo y las 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.

- **Revisar el proceso.**
  Este proceso se refiere a calificar al proceso entero de minería de datos a objeto de identificar elementos que pudieran ser mejorados.

- **Determinar los próximos pasos.**
  Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios podría pasarse a la siguiente fase, en caso contrario podría decidirse por hacer otra iteración desde la fase de preparación de los datos o de modelado con distintos parámetros. Podría incluso darse el caso de que en esta fase se decida empezar desde cero con un nuevo proyecto de minería de datos.

**6. Despliegue e Implantación.**

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o por ejemplo aplicando el modelo a diferentes conjuntos de datos o como parte del proceso (en análisis de riesgo de créditos, detección de fraudes, etc.). Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que 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 componen esta fase (figura 10) son:




- **Planear la implantación.**
  Para implementar el resultado de la minería de datos 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 Aplicación de la Metodología CRISP-DM

a un Proyecto de Minería de Datos en el Entorno Universitario - 33 - modelo, este procedimiento debe ser documentado para su posterior implementación.

- **Planear la monitorización y mantenimiento.**
  Si los modelos resultantes del proceso de minería de datos 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.

- **Producir el informe final.**
  Es la conclusión del proyecto de minería de datos realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia adquirida o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

- **Revisar el proyecto.**
  En esta tarea se evalúa que cosas se hicieron correctamente y cuales fueron incorrectas, así como aquellos puntos que se podrían mejorar en el proyecto.