UNIVERSIDAD
DE LOS ANDES
DR. PEDRO RINCÓN GUTIÉRREZ
TACHIRA VENEZUELA

Metodología para identificar donde extraer conocimiento en una organización

Rangel, C. Pacheco, F.,Aguilar, J., Cerrada, M., Altamiranda, J.
Universidad de los Andes
{crrp88,fannikaro}@gmail.com ,{aguilar,cerradam,altamira}@ula.ve

Enero, 2013.

Proyecto:

Desarrollo de herramientas computacionales basadas en técnicas inteligentes para la gestión de bases de datos sobre las actividades nacionales en salud y petróleo, para realizar tareas de minería de datos.

Grupo de Investigación:

- Coordinador: José Aguilar
- Responsables: Mariela Cerrada
  Junior Altamiranda
- Investigadores: Fannia Pacheco
  Carlos Rangel

Problema:

Actualmente las organizaciones poseen grandes cantidades de datos que no son utilizados eficientemente. Desde dichos datos, entre otras cosas, se pueden extraer conocimientos útiles para dichas organizaciones en sus procesos de toma de decisión. Esos conocimientos pueden ser usados en tareas de predicción, clasificación, optimización, etc.

Objetivo

- Desarrollo de herramientas computacionales basadas en técnicas inteligentes (Redes Neuronales, Computación Evolutiva, Lógica Difusa, etc.) para la extracción de conocimiento desde bases de datos del sector salud y petrolero

Resultados Esperados

- Componentes de software especializados en tareas de minería de datos para el sector petrolero y de salud nacional
- Formación de recurso humano especializado en técnicas inteligentes
- Optimización de procesos de tomas de decisión organizacional nacional

Para el desarrollo del presente proyecto, es necesaria una fase de ingeniería de conocimiento, la misma permitirá al grupo de investigación un amplio conocimiento del proceso a estudiar. La Ingeniería del Conocimiento involucra una variedad de personas:

- El(los) ingeniero(s) de conocimiento, que es la persona encargada de la construcción y puesta en marcha del proyecto, en este caso es el grupo de investigación.
- Un experto o grupo de expertos en el dominio, vinculado(s) a las instituciones con las que se articulará el proyecto.

Se plantea una metodología para identificar donde extraer conocimiento, para una adecuada interacción entre el grupo de investigación y los expertos en el dominio.

Metodología para el desarrollo de aplicaciones de minería de datos, basada en el análisis organizacional

Esta metodología está diseñada para para el desarrollo de aplicaciones basadas en MD para un proceso de cualquier institución/empresa. Está compuesta por tres grandes fases, como lo son: (i) Identificación de fuentes para la extracción de conocimiento en una organización, (ii) Preparación y tratamiento de los Datos y (iii) Desarrollo de herramientas de MD. En la Figura 1 el flujo de desarrollo de dichas fases, se observa que se puede retroceder a fases anteriores de ser necesario

Figura 1: Flujo de desarrollo de la metodología propuesta.

Cada fase de la metodología aquí propuesta está concebida en etapas que se ejecutan secuencialmente mediante una serie de pasos. Los elementos que conforman cada etapa se muestran en la Figura 2.

Figura 2: Aspectos que conforman cada etapa de las fases de la metodología.

Así, en cada una de las etapas se especifican esos cuatro elementos, para que el proceso de ingeniería de conocimiento avance de manera adecuada:

1. Objetivo: describe cual es la meta que se quiere cumplir en la etapa respectiva
2. Producto principal: que es lo que se debe producir, en concreto, al final de la etapa.
3. Protocolo: describe los elementos que se deben investigar ó conocer en la etapa. En general, un protocolo es el conjunto de procedimientos, preguntas o estudios que se deben realizar para desarrollar la etapa.
4. Actividades: describe las tareas que se designan a los investigadores y a la organización/empresa lograr el objetivo de la etapa.

1. Fase 1: Identificación de fuentes para la extracción de conocimiento en una organización

Esta fase tiene como finalidad realizar un proceso de ingeniería de conocimiento, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s), o los procesos a estudiar. Esta etapa se enfoca a identificar y conceptualizar la solución de un problema, desde la perspectiva del desarrollo de aplicaciones basadas en MD.

El principal objetivo de esta fase es conocer la organización, sus procesos, sus expertos, entre otros aspectos, para definir el objetivo de la aplicación de la MD en la organización, mediante el uso de preguntas, actividades estructuradas y documentos. En la Figura 3 se observan los pasos que conforman esta fase, recordando que cada paso se define como una etapa y cada etapa tiene: objetivos, producto principal, protocolo y actividades.

Figura 3: Etapas que conforman la fase 1.

1.1. Conocimiento de la organización

a) Objetivo:

El objetivo de esta etapa es conocer la organización/empresa, sus objetivos, procesos, objetos y actores, para ello se requiere de una breve y consistente información sobre la historia, objetivos y organización de la institución/empresa por parte de los expertos, para que los ingenieros de conocimiento se familiaricen con los propósitos de la organización.

b) Producto principal

Un documento con toda la información que permita conocer la institución/empresa, o documentos equivalentes.

El documento contiene por lo menos los siguientes ítems:

- Descripción de los elementos de la institución/empresa y sus características
- Descripción de las relaciones entre estos elementos
- Organización de estos elementos

c) Protocolo

Hay diferentes elementos presentes en una organización, se consideran como más importantes, los siguientes:

- Objetivos
- Procesos
- Objetos
- Actores

Para la descripción de cada elemento, se pueden realizar las preguntas dadas en la Tabla 1.1.

Tabla 1.1: Preguntas y ejemplos para determinar los elementos de la institución/empresa

Elemento	Preguntas	Ejemplos
Objetivos	¿Cuál es la razón de ser de la institución?	Conocer, determinar, establecer, la finalidad de la institución/empresa.
Procesos	¿Cuales son las actividades que permiten alcanzar los objetivos de la institución?	Procesos de producción o administrativos.
Objetos	¿Qué cosas o entidades se manipulan en los procesos de la institución?	Pueden ser físicos o abstractos, departamentos, documentos, herramientas, plantas.
Actores	¿Quiénes ejecutan los procesos?	Personas, sistemas, máquinas, etc.

d) Actividades:

- Por parte de la institución/empresa:

Actividad: Generar un documento que permita conocer la institución/empresa, respondiendo las interrogantes de la tabla 1.1, para proveérselo al grupo de investigación. En caso de tener un documento equivalente, facilitárselo a los investigadores (por ejemplo: documento organizacional, organigrama de la institución/empresa, etc.).

Este documento deberá ser consignado lo antes posible, para que cuando éstos visiten la institución/empresa tengan conocimiento previo de los elementos importantes en ella.

Momento: Primera actividad que realiza la institución/empresa, previa a la primera visita.

- Por parte de los ingenieros de conocimiento:

Actividad: Estudio de la institución/empresa con la información proporcionada por la misma. Generar dudas sobre el funcionamiento de la institución/empresa, de sus procesos y objetivos expuestos en el documento consignado al grupo de investigación.

Momento: Una vez consignado el documento por la institución/empresa, previo a la primera visita.

- Trabajo conjunto:

Actividad: Planificación de la primera entrevista, para que el grupo de ingenieros de conocimiento aclare las dudas que tiene acerca de la organización, conozca mejor los procesos descritos en el documento, etc. El grupo de ingenieros de conocimiento deberá solicitar entrevistas con ciertos actores en los procesos de interés (expertos en los procesos), así como también con otros actores en la parte administrativa o gerencial, si son pertinentes en el proceso en cuestión.

Momento: Durante la primera visita.

1.2. Caracterización detallada de los procesos de la organización

a) Objetivo:

Esta etapa tiene como finalidad conocer en detalle los procesos sobre los cuales se puede enfocar el proyecto de minería de datos, para ello se formulan un conjunto de preguntas que servirán de apoyo para el desarrollo de esta etapa.

b) Producto principal

Documento que contiene el flujo de los procesos de la organización, modelos de procesos y diagramas de actividades

c) Protocolo

Esta etapa es realizada por los expertos de la organización y se desglosa en los siguientes pasos:

1.2.1 Familiarización con los procesos sobre los cuales se puede realizar la extracción de conocimiento

- ¿Qué productos generan esos procesos?
- ¿Qué beneficios proporcionan esos procesos a la organización?
- ¿Qué problemas tienen actualmente?
- ¿Importancia de esos procesos para la organización, o impacto sobre otros procesos?
- ¿Qué impacto generaría la mejora de esos procesos o el estudio de los mismos?

1.2.2. Identificar la fuente del conocimiento

- ¿Cuáles son los actores o personas que intervienen en los procesos?
- ¿Quién o quiénes son las personas expertas en los procesos?
- ¿Existen documentos que permitan conocer esos procesos?
- ¿Existen sistemas computacionales que intervengan o interactúen en el proceso?

1.2.3. Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso explicado

- ¿Dónde se encuentra los datos almacenados del proceso en cuestión?
- ¿Cómo se almacenan los datos del proceso?
- ¿Qué variables son observadas del proceso?
- ¿Cuáles son las variables más importancia de esos datos para la organización?

d) Actividades

- Por parte de la institución/empresa:

Actividad: Generar un documento y/o presentación con el(los) proceso(s), que conteste las preguntas del punto 2 para proveérselo a los investigadores. En caso de tener documentos equivalentes, facilitárselos al grupo de investigación.

Momento: Previo a la primera visita.

- Por parte de los ingenieros de conocimiento:

  Actividad: Estudio de los procesos con la información proporcionada por la organización. Generar un cuestionario sobre las dudas que se tengan acerca de los procesos.

  Momento: Previo a la primera visita.
- Trabajo conjunto:

  Actividad: Entrevista para aclarar las dudas y preguntas, que generó el grupo de ingenieros de conocimiento.

  Momento: Durante la primera visita.

1.3. Análisis de factibilidad y selección del proceso

a) Objetivo:

En esta etapa se requiere un análisis de cada proceso estudiado en el paso anterior, con la finalidad de conocer la factibilidad de la aplicación de la minera de datos sobre cada uno de ellos. Para ello se utilizan criterios que permitirán finalmente la selección de uno o más procesos que cumplan con las características necesarias para la aplicación de la tarea de minería de datos.

b) Producto principal:

Tabla con evaluación ponderada de parámetros de selección de proceso de interés de la organización.

c) Protocolo:

Con la información proporcionada/recogida (pasos 1.1 y 1.2), donde se expresan los procesos de interés, deberá hacerse una selección de cuáles de estos procesos son viables para tratarse usando minería de datos. Este estudio lo realizan los ingenieros de conocimiento, tomando en cuenta los siguientes aspectos:

1.3.1. Revisión de los procesos propuestos por los expertos

- Revisión de la literatura existente acerca de problemas semejantes, que se hallan tratado con minería de datos

- Análisis detallado de los documentos proporcionados por la institución/empresa
- Determinación del propósito de aplicar minería de datos en los procesos

1.3.2. Importancia de los procesos para la organización

- Revisión del documento proporcionado por la institución/empresa, para observar la importancia que tienen esos procesos
- Búsqueda de ejemplos, donde se hayan obtenido resultados satisfactorios en problemas semejantes

1.3.3. Disponibilidad del experto o grupo de expertos

- Verificar por medio de la entrevista realizada por el grupo de ingenieros de conocimiento, cuál es la disponibilidad de atención de los expertos.

1.3.4. Análisis de las fuentes de información sobre los procesos

- Con los documentos proporcionados, verificar si las fuentes de información son tratables para la aplicación de minería de datos.
- Disponibilidad de datos y herramientas computacionales con las que se puedan manejar.
- Observar el historial de datos que se almacena
- Verificar si los datos son representativos para realizar minería de datos
- Verificar los sistemas computacionales existentes a nivel de: su operatividad, etc.

Para la selección del proceso(s) a considerar para realizar la(s) tarea(s) de MD se usan los criterios descritos en la Tabla 1, que son aspectos estudiados en el protocolo de esta etapa.

Tabla 1. Criterios para la selección del(los) proceso(s)

Criterios	**Descripción**
Importancia para la institución/organización	Nivel de importancia que la organización le tiene al proceso, basándose en una numeración del 1 al 5, donde el 5 es el más importante.
Propósito de la MD	Impacto que generaría mejorar este proceso usando MD
Interacciones entre procesos	Cantidad de interacciones que posee el proceso con otros procesos de interés.
Procesos dependientes	Cantidad de procesos que dependen del proceso en cuestión.
Importancia de la calidad del producto	Basándose en una numeración del 1 al 5 donde el 5 es el más importante. Se mide que tan importante es el producto que se obtiene del proceso estudiado sea de calidad.
Seguridad Industrial	Describe si el proceso en cuestión es de alto riesgo en factores de seguridad industrial. Los valores serán tomados como el 1 el de menor riesgo y 5 el de mayor riesgo. Para el total ponderado de las priorizaciones este valor restará (será negativo en la suma) peso.
Replicabilidad de la herramienta desarrollada	Si escogiendo este proceso la herramienta puede o no ser aplicada a otras organizaciones de índole similar. Siendo 1 el valor menos importante y 5 el valor más importante.
Cantidad de Expertos	Cantidad de expertos en el área relacionada al proceso en cuestión.
Fuentes de información	Calidad de la fuente de información, medida con una numeración del 1 al 5 donde 5 es excelente
Confidencialidad de la información	Si los datos tratados son de poca o alta confidencialidad. Los valores serán tomados como el 1 el de menor confidencialidad y 5 el de mayor confidencialidad. Para el total ponderado de las priorizaciones este valor restará (será negativo en la suma) peso.
Que información se recoge del proceso para	Cantidad de información que recoge el proceso.

ser almacenada	
Con que frecuencia se recoge la información almacenada	Frecuencia en que se toma la información almacenada para este proceso. Medida con una numeración del 1 al 5 donde 5 es excelente
Que herramientas se cuentan, para recolectar y manipular la información	Cantidad de herramientas que cuenta la organización para recolectar y manipular la información.

Para los criterios cualitativos, se toman los valores numéricos que miden su importancia en la organización.

Cuando no se tiene información de algún criterio, ya sea porque no se tienen en la institución/empresa, o no son relevantes para ella, se dejan en cero.

Para la selección del proceso, se sustituyen los valores en la tabla 1 de cada proceso de la institución/empresa y se realiza la suma aritmética de los valores numéricos, seguidamente se realiza una suma ponderada, con los pesos previamente fijados según nivel de importancia. En la ecuación (1) se describe la suma ponderada, donde T_k es la suma ponderada para cada proceso k, C_i es un criterio con ponderación p_i y n es la cantidad de criterios definidos en la tabla V. El proceso que de cómo resultado la suma ponderada más alta (T_k) será el seleccionado.

$$T_k = \sum_{i}^{n} C_i * p_i \qquad (1)$$

Los campos Seguridad Industrial y Confidencialidad de la información, los cuales servirán para ayudar a escoger el(los) proceso(s) a ser estudiado(s), restan valor en las sumas, ya que son criterios que pueden retrasar la tarea de minería de datos.

d) Actividades:

- Por parte de los ingenieros de conocimiento:

  Actividades: Estudio de los procesos con la información proporcionada por la misma. Analizar los procesos dados por la institución/empresa.

  Momento: Previo a la segunda visita.

- Trabajo en conjunto:

  Actividad: La institución/empresa solventará dudas surgidas por el grupo de investigación para ayudar con la selección del proceso.

  Momento: Durante el desarrollo de esta fase, vía virtual.

Una vez conocida la empresa/institución, recabada información de la misma, analizados sus procesos, y seleccionado el proceso sobre el cual se realizará la tarea MD por su importancia para la empresa/institución, se procede a caracterizar las posibles tareas de MD a realizar en dicho proceso (objetivos, requerimientos, factibilidad, etc.), con la finalidad de escoger la tarea de minería de datos especifica a desarrollar.

1.4. Análisis para caracterizar las posibles tareas de Minería de Datos (MD)

a) Objetivo:

En esta etapa se desea caracterizar las posibles tareas de MD a realizar en el(los) proceso(s) seleccionado(s) en la fase anterior (objetivos, requerimientos, factibilidad, etc.), con la finalidad de escoger las tareas de MD de interés a desarrollar.

b) Producto Principal:

Documento de requisitos funcionales, casos de uso, actores involucrados en el proceso estudiado, tablas de escenarios actuales y futuros para aplicar MD.

c) Protocolo

Los ingenieros de conocimiento realizarán una serie de preguntas a personas adecuadas en la institución/empresa para así obtener las necesidades de la institución, acorde al proceso escogido previamente. Dichas preguntas se pueden realizar mediante entrevistas individuales a expertos en el proceso, o entrevistas grupales a grupos de expertos en el proceso en cuestión.

Para caracterizar las posibles tareas de MD se usará la idea de *escenarios*. Entenderemos por escenario una descripción de un resultado, los actores involucrados para obtener dicho escenario, las variables asociadas, y actividades que se realizan para llegar al resultado. Los escenarios pueden ser el actual, el cual

es una descripción del comportamiento actual del sistema, que permite conocer cómo se están obteniendo los resultados de los procesos, y futuros en el cual se da una descripción general de los resultados esperados o deseados que se pueden obtener después de aplicar la tarea de MD al escenario actual.
Así, en esta etapa se realizan los siguientes pasos:

1.3.5. Selección y descripción de los actores.

Tomando en cuenta las definiciones y especificaciones hechas en los puntos anteriores, se seleccionan los actores involucrados en el proceso con los que se trabajarán. Dichos actores pueden ser equipos o humanos, siendo un especial tipo de actor los expertos en los procesos, quienes conocen el funcionamiento y las actividades de los mismos. Algunas preguntas que ayudan a describir a los actores de un proceso son:

- ¿Qué tareas desempeña cada actor en el proceso?
- ¿Qué información requiere cada actor para cumplir las tareas que desempeñan?
- ¿De cuáles eventos e información sobre el proceso son informados los actores del proceso?
- ¿Existe interacción entre los actores? De haber interacción, describirla.
- ¿Qué información o tareas comparten los actores?
- ¿Qué cambios en los procesos deben ser informados a/por los actores?
- ¿Qué actividades se realizan al ocurrir los cambios planteados en la pregunta anterior?
- ¿Qué funcionalidades no tienen los actores en este momento, pero que pudieran tener?

1.3.6. Descripción de los escenarios.

Determinar los escenarios del proceso por medio de entrevistas, usando una serie de preguntas generadas por el grupo de ingenieros de conocimiento hacia los expertos. Para describir a los escenarios, se define la noción de

variable como el elemento que caracteriza algún aspecto del proceso que puede variar en el tiempo.

Preguntas a los expertos para caracterizar a las variables de un proceso:

- ¿Cuál es el flujo de actividades detallado del proceso en estudio? (De existir un diagrama de actividades continuar con las preguntas) Para ayudar a desarrollar el diagrama de actividades, se deben contestar las siguientes preguntas:
    - ✓ ¿Cuáles son las variables más importantes observadas en el proceso estudiado?
    - ✓ ¿Cuáles de estas variables son críticas para la toma de decisiones del proceso?
    - ✓ ¿Cuáles de estas variables son críticas para la toma de decisiones de otros procesos?
    - ✓ ¿Qué interacciones existen entre las variables? (de existir)
- Al observar dichas variables
    - ✓ ¿Qué se conoce del resultado global del proceso?
    - ✓ ¿Qué se podría inferir del resultado global del proceso?
    - ✓ ¿Cómo afecta al resultado global del proceso?
    - ✓ ¿Cómo afecta el resultado de este proceso a otros procesos?
    - ✓ ¿Qué otra información puede extraerse de estas variables? (si tienen conocimiento de ello)
- ¿Dichas variables pueden modificarse al haber algún cambio en el proceso asociado? ¿Es factible inducir cambios el proceso? ¿Cómo se pueden inducir esos cambios en el proceso?

- Descripción detallada del escenario actual con ayuda de los expertos, para ello es necesario completar la Tabla 1, en la cual se describe cual(es) es(son) el(los) resultado(s) que se obtiene(n) en el escenario actual asociado a un proceso, los actores involucrado, las variables asociadas y las actividades que se siguen para obtener el(los) producto(s)

Tabla 1. Estructura de la tabla que permitirá describir un escenario actual.

Resultados que se obtienen	Actor(es) asociado(s)	Variables Asociadas	Actividades que se realizan
Producto(s)	Actor(es) que interviene(n) para el desarrollo del producto	Variables que están relacionadas con el producto	Actividades que se realizan para obtener el producto

- A partir del escenario actual, los ingenieros de conocimiento definirán los escenarios posibles o hipotéticos, relacionados con las tareas de MD posibles a aplicar. Este proceso de prospectiva tecnológica que se realiza, permite a la organización definir funcionalidades con relevancia, que en un futuro desean obtener basándose en el estado actual que se encuentra. Para ello se usará la siguiente tabla (se realizará un escenario por cada posible tarea de MD a aplicar):

Tabla 2. Estructura de la tabla que permitirá describir los escenarios futuros.

Resultados que se desean obtener	Actor(es) asociado(s)	Variables Asociadas	Actividades de MD que se realizarían	Funcionalidades nuevas
Producto deseado que se pueden obtener por medio de MD	Actor(es) que interviene(n) para el desarrollo del producto	Variables que están relacionadas con el producto	Actividades que se realizan para obtener el producto	Funcionalidades que no tiene el actor(es)/actividades/proceso en este momento, pero que pudieran tener

- Selección de los escenarios factibles de MD. A partir de esa selección, se concibe el(los) *escenario(s) futuro(s)* (puede ser uno de los factibles, varios escenarios, la fusión de algunos). Sin embargo los demás escenarios futuros no son descartados, ya que es posible que sean estudiados más adelante en otros proyectos. El conjunto de escenarios futuros que no son escogidos para el proyecto en desarrollo, queda como insumo a la organización como un plan tecnológico, producto de la prospectiva

tecnológica realizada. Este paso se realiza en una reunión entre el grupo de expertos y el grupo de ingenieros de conocimiento. Para escoger los escenarios factibles se usan los siguientes criterios:

Tabla 3.Criterios para selección del escenario futuro.

Criterios	Descripción
Importancia del resultado que se espera del escenario para la empresa/institución	Nivel de importancia del escenario propuesto, basándose en una numeración del 1 al 5 donde el 5 es el más importante.
Utilidad del escenario para la empresa/institución	Utilidad del escenario futuro, basándose en una numeración del 1 al 5 donde el 5 es el más útil.
Cantidad de expertos asociados al escenario	Cantidad de expertos en el área relacionada al escenario en cuestión.
Seguridad Industrial (si aplica)	Basándose en una numeración del 1 al 5 donde el 5 es el más alto. Se mide que tan importante es la seguridad industrial en el escenario.
Fuentes de información requeridas por el escenario	Calidad de la fuente de información, medida con una numeración del 1 al 5 donde 5 es excelente.
Confidencialidad de la información	Confidencialidad de la información para la empresa, lo que permitirá o no proveerla a los investigadores para el desarrollo del escenario futuro
¿Con que frecuencia se recogen los datos almacenados asociados a la información de interés?	Frecuencia con que se toma la información almacenada para este proceso. Medida con una numeración del 1 al 5 donde 5 es excelente
¿Con qué herramientas se cuenta para recolectar y manipular los datos?	Cantidad de herramientas que cuenta la organización para recolectar y manipular la información.
Replicabilidad de la herramienta a desarrollar en otros escenarios	Uso de la aplicación desarrollada en otras empresas que estén compuestas por procesos semejantes, o en otros procesos de la empresa

1.3.7. Especificación de los requerimientos para el plan tecnológico de desarrollo del(los) escenario(s) futuro(s) (tarea(s) de MD a aplicar)

Para cada uno de los escenarios futuros definidos y seleccionados en la etapa anterior, se definen un conjunto de requerimientos funcionales y no funcionales. Esa tarea es realizada por los ingenieros de conocimiento.

- Requerimientos Funcionales
  - ¿Qué funciones debe cumplir la tarea de MD en el escenario escogido?

- ¿Qué interacción tendrá la tarea de MD de datos con los actores del escenario?
- ¿Qué interacción tendrá la tarea de MD de datos con el escenario actual del proceso escogido?
- ¿Qué interacción tendrá la tarea de MD de datos con otros escenarios actuales de otros procesos?

- Requerimientos no Funcionales
  - ¿En cuál plataforma el sistema debe ser implementado?
  - ¿Qué características debe cumplir la implementación de la tarea de MD en la plataforma?
  - Identificar los datos de entrada para la tarea de MD, y las herramientas/sistemas con que cuenta la organización para proveerlos (pueden ser datos de entrada o salida del proceso)

Tabla 4. Tabla para describir los requerimientos funcionales.

Id del requerimiento: F# para requerimientos funcionales y N# para requerimientos no funcionales, donde # es un número incremental partiendo desde el 0.	**Prioridad:** Etiqueta que describe la prioridad del requerimiento que puede ser: Alta, Media, Baja
Nombre del requerimiento:	Nombre que identifique el requerimiento.
Descripción del requerimiento:	Descripción detallada de las características que se desean alcanzar con dicho requerimiento.
Escenario(s) asociado(s):	Interacción que tendrá el requerimiento con el escenario actual del proceso escogido u otros procesos
Actores asociados:	Interacción que tendrá el requerimiento con los actores del escenario
Id(s) de los Requerimientos	De existir requerimientos relacionados con el mismo, colocar el(los) Id(s) del(los) requerimiento(s).

asociados:	

1.3.8. Elaboración de los casos de uso para los requerimientos funcionales

- Generar los casos de uso que solventarán los requerimientos funcionales especificados. Esta tarea la realizan los investigadores. Dichos casos de uso serán diagramados usando UML.

1.3.9. Determinación del alcance de la investigación

- Elaborar el plan preliminar de actividades para el desarrollo de herramienta de MD.

d) Actividades

- Por parte de los ingenieros de conocimiento:

  Actividad: Generar un documento de requisitos, casos de uso para estos requisitos, actores involucrados en el proceso estudiado para aplicar minería de datos.

  Momento: Previo a la segunda visita.

- Trabajo conjunto:

  Actividad: Refinar el documento generado por parte de los ingenieros de conocimiento. Dicho documento servirá como compromiso preliminar de las metas a cumplir (requisitos), teniendo en cuenta que este mismo puede cambiar a medida que se conozca mejor el proceso estudiado y la data almacenada de éste.

  Momento: Durante la segunda visita.

1.5. Formalización de las tareas de Minería de Datos (MD)

a) Objetivo

Definir el(los) problema(s) formales de MD.

b) Producto principal

Documento formal con la definición del problema.

c) Protocolo de etapa

Desarrollo de un informe por parte del grupo de investigación, con la conceptualización del proceso a estudiar, la caracterización de sus problemáticas operacionales y del uso de la MD en dicho proceso. Además, ese documento debe contener el cronograma de la fase de MD, según la metodología que se use. Este documento es una versión inicial de la definición del problema, que irá evolucionando a medida que el proyecto avance.

d) Actividades

- Por parte de los investigadores:

  Actividad: Generar un documento formal con la definición del problema.

  Momento: Después de la segunda visita.

2. Fase 2: Preparación y tratamiento de los Datos

Para aplicar MD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema de estudio. Esto conlleva realizar distintas operaciones con los datos, con la finalidad de acondicionarlos para desarrollar un modelo de MD. Para realizar este proceso se crean diferentes vista minables, que básicamente contienen la información de las variables y los datos del historial, a continuación se definen con más detalle:

Una vista minable conceptual describe en detalle cada una de las variables a ser tomadas en cuenta para la tarea de Minería de Datos, para cada escenario futuro seleccionado. La misma está compuesta por todas las variables de interés, y algunos campos adicionales de importancia para realizar el proceso de tratamiento de datos (como por ejemplo: dependencias con otras variables, redundancia de medición, entre otras características que se consideren importante).

Por otro lado, para realizar tareas de MD es necesario tener los datos integrados en una sola vista, para esto se ha venido definiendo la vista minable conceptual. Ahora, a esta vista minable conceptual es necesario empezar a cargarla con datos, esta misma al ser cargada con datos, la llamaremos vista minable operativa.

Por lo tanto se plantea esta fase para realizar la preparación y tratamiento adecuado de los datos, que serán utilizados para el desarrollo de la herramienta de MD. En la Figura 4 se muestran las etapas que conforman esta fases.

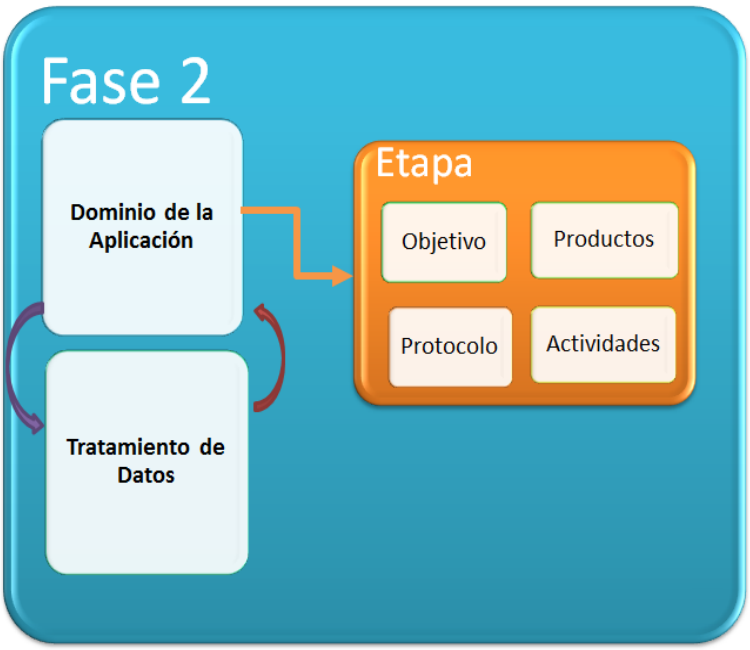

Figura 4: Etapas que conforman la fase 2.

2.1. Dominio de la aplicación

a) Objetivos

En esta etapa se deben producir dos aspectos concretos, la vista minable conceptual y la vista minable operativa.

Otros objetivos serian:

- Ubicar y comprender los datos asociados a el(los) escenario(s) futuro(s)
- Construccion de una vista minable conceptual que tiene las variables de interes para en caso de estudio
- Construccion de una vista minable operativa que esta compuesta por datos
- Definicion de la(s) variable(s) objetivo(s) en la vista minable operativa

b) Productos principales

- Características de los repositorios donde se encuentran los datos
- Vista minable conceptual
- Vista minable operativa
- Descripción de la(s) variable(s) objetivo(s)

c) Protocolo de etapa

2.1.1. Comprensión de los datos de entrada

Según el escenario futuro sobre el cual se esté realizando el estudio, es importante tener conocimiento de los siguientes aspectos:

a. Comprensión de los datos asociados a las variables

   Explicar que se entiende por datos asociados a la variable: unidades, tipos, etc.

   - ¿Cuáles son estos datos?
   - ¿Cuáles son las características de esos datos? Por ejemplo: restricciones, rangos de medición, unidades, etc.

b. Determinación de los repositorios de datos

   - Tipos de archivos en la que se almacena los datos (los cuales pueden ser físicos o digitales)
   - Organización de la base de datos (en caso que existan datos llevados de manera manual, estos deben ser digitalizados para su futuro tratamiento)
   - ¿Errores comunes en la adquisición de estos datos?
   - Otras anomalías

2.1.2. Construcción de la vista minable conceptual

En este paso se definen cada una de las variables de manera detallada asociadas al proceso de interés, mediante el uso de una vista minable conceptual. Los pasos para definir dicha vista se muestran a continuación:

- Realizar un primer filtrado, en este paso es necesario seleccionar las variables de interés para el escenario en estudio, dicho filtrado se realiza con los expertos del proceso y los ingenieros de conocimiento.
- Establecer las relaciones entre las variables seleccionadas (dependencia entre variables, redundancia, variables que son producto de formulas entre otras variables), se establecen los campos adicionales previamente mencionados,

se colocaran tantos campos como sean necesarios, ya que esto ayudara al proceso de tratamiento de datos y hasta a la propia tarea de minería de datos.

- Extender la vista minable conceptual en base a las necesidades de los escenarios (de ser necesario): estudiando el escenario futuro, observar si es necesario extender la vista minable conceptual con otras variables que puedan aportar información (variables de otros procesos que puedan estar influyendo en el proceso, pero que en la actualidad no son tomadas en cuenta); dicha extensión depende del conocimiento adicional que pueda aportar el experto.

2.1.3. Integración de los datos de entrada

Una vez obtenida la vista minable conceptual, se procede a cargarla con datos del historial, convirtiéndose en una vista minable operativa.

En esta fase, si los datos están en repositorios distintos, lugares distintos, o que por estar en lugares diferentes se llaman diferente, sencillamente se deben integrar para esto se debe Tipificar la integración que se realizará: ver datos comunes, fusionarlos, que tipo de dato va a quedar, que nombres van a quedar., e integrar formatos.

Así, todos los datos que la aplicación manejará deben estar en un mismo repositorio. La integración de estos datos debe darse en un repositorio (físico o digital), que los ingenieros de conocimiento tengan libre acceso. Esta integración dará como resultado la vista minable operativa, dicha vista es una tabla donde se encuentran todos los datos a manipular.

Se deben realizar los siguientes pasos, de ser necesaria una integración de datos:

- Si se encuentran en diferentes repositorios ubicarlos
- Observar la organización en la que están dispuestos los datos en cada repositorio y como se almacenan
- Definir una estrategia para unificar los datos en un solo repositorio.
- Integrar formatos.

- Crear la vista minable operativa, resultante de la integración de los datos asociados a las variables escogidas en la vista minable conceptual (fusión de tablas, integración de bases de datos, entre otros).

2.1.4. Definir las variables objetivos

Una vez planteado el escenario futuro y la tarea de minería de datos a realizar, es preciso detectar las variables que permitirán la consecución de los objetivos de MD, a estas variables se le denominan variables objetivos, ya que las mismas son las que se desean desea predecir, clasificar, calcular, inferir, en otras palabras, es la que deseamos obtener con la tarea de MD. Así, en esta fase se desea definir las variables objetivo del escenario futuro seleccionado y ubicar dichas variables en la vista minable descrita previamente.

Además, una vez planteado el escenario futuro y la tarea de minería de datos a realizar, es preciso detectar las variables que aportaran en gran medida que la tarea de MD se desarrolle de manera correcta (variables significativas). Determinar de todas las variables descritas en la vista minable operativa la(s) variable(s) de interés (variables objetivos y significativas), será es el objetivo de este punto. Para ello se deben realizar los siguientes puntos:

- Teniendo en cuenta las entradas, ¿a qué conclusiones puede llegar el experto humano?
- Observar el objetivo en el escenario futuro seleccionado e identificar ¿Cuál de las variables llevan a dicho objetivo?
- Escoger la(s) variable(s) objetivo(s)

d) Actividades

- Por parte de la institución/empresa:
  - Proporcionar a los investigadores la información de los datos asociado al escenario futuro seleccionado

- Proveer los datos asociados a la vista minable conceptual provenientes de los servidores de la institución/ empresa
- Proporcionar a los investigadores información que les permita definir las variables objetivo

- Por parte de los ingenieros de conocimiento:
  - Generar una descripción de los datos y las relaciones que tienen con las variables.
  - Conocer la como están almacenados los datos
  - Construir la vista minable operativa
  - Seleccionar y ubicar la(s) variable(s) objetivo(s) en la vista minable con datos
- Trabajo conjunto:

  Reuniones virtuales para completar los puntos a, b y c.

2.2. Tratamiento de datos

a) Objetivos

Esta etapa se centra en generar datos de calidad, es decir datos sin anomalías, sin inconsistencias de formato, sin capturas erróneas, sin campos vacíos; aplicando metodos de limpieza, transformación y reducción sobre la vista minable operativa.

b) Productos principales

- Vista minable tratada

c) Protocolo de etapa

La vista minable operativa es preparada mediante herramientas especializadas para realizar limpieza de datos innecesarios, transformación de las variables observadas, reducción de variables, entre otros métodos que se requieran para generar una vista minable operativa de calidad. Cabe destacar que ya existen diferentes técnicas y

algoritmos para realizar esta etapa (como lo es el análisis de correlación y el cálculo de la entropía).

El tratamiento de datos se va aplicar sobre la vista minable operativa, la cual se manipulara siguiendo los siguientes pasos:

2.2.1. Limpieza

La limpieza de datos se refiere a una serie de procesos en los cuales la calidad de los datos es mejorada, enfrentando los problemas mencionados como datos mal capturados, anómalos y vacíos, ya sea por características obvias que el dato no cumple con ciertos parámetros del estándar, o porque el experto del proceso ya tiene identificado anomalías comunes en el almacenamiento de los datos.

En general, en el proceso de limpieza realiza normalización de formatos, remoción de anomalías, corrección de errores y eliminación de duplicados. Una técnica que se puede mencionar es limpiar datos anómalos que se alejen mucho de la media estándar de los datos, ya que estos datos describen sucesos tales como: fueron mal tomados, se almacenaron de forma incorrecta o que son simplemente una instancia que sí ocurrió pero es poco probable que vuelva a ocurrir. Este tipo de datos pueden generar cierto ruido en el estudio y por eso es mejor eliminarlos.

En esta etapa se deben buscar las anomalías que presenta la base de datos, tales como:

- Unidades de las entradas
- Abreviaciones
- Convenciones de nombres
- Representaciones diferentes
- Variaciones de Ortografía
- Elementos repetidos
- Datos no guardados

Para identificar las anomalías que se están buscando se debe:

- Estudiar la representación de cada una de las variables.
- Buscar anomalías de representación.
- Después de buscar las anomalías presentes en la base de datos, definir alguna estrategia de limpieza para erradicar dichas anomalías y obtener data consistente.
- De acuerdo a la representación de las variables, realizar las operaciones con un software para limpieza de datos.

2.2.2. Transformación

Las transformaciones consisten principalmente en modificaciones sintácticas llevadas a cabo sobre la vista minable operativa, que no impliquen un cambio en el significado de los mismos, y además, que sea conveniente a la hora de aplicar MD. El uso de esta etapa depende del problema y del investigador que realiza el diseño de la herramienta de MD.

En esta etapa se transforma variables de entrada en nuevas variables de interés, esto se realiza a través de diversos métodos, los cuales se deben escoger en caso de ser pertinente alguna transformación de alguna de las variables. Una transformación de variables puede ser la combinación entre variables (concatenación de cadenas, multiplicación entre variables, entre otras operaciones aritméticas).

- Estudiar las representaciones de cada una de las variables
- Identificar las representaciones que se puedan transformar en otra representación más conveniente o fácil de utilizar a la hora de aplicar la tarea de minería de datos.
- Ordenar dichas transformaciones que se desean aplicar en una tabla, para observar las equivalencias
- Aplicar la transformación con el software seleccionado
- Identificar las variables que potencialmente se pueden normalizar
- Definir la función(es) de normalización para cada una de las variables seleccionadas en el paso anterior y ordenarla en tablas.
- Aplicar la función(es) de normalización en las variables seleccionada

- De ser necesario, combinar variables por un método seleccionado tal como el PCA (del inglés *Principal Component Analysis*) que es considerado también un método para reducción de variables.
- Describir en tablas cada una de las transformaciones realizadas.

2.2.3. Reducción

Consiste en decidir qué datos deben ser utilizados para el análisis. El criterio que se sigue para realizar reducción de variables presentes en la vista minable operativa incluye la relevancia con respecto a los objetivos que se persiguen en la MD, y limitaciones técnicas tales como los volúmenes máximos de datos o bien tipos de datos concretos. En esta fase nos centraremos en el último caso, ya que el primero fue abarcado en los pasos anteriores (2.1.2-2.1.3).

Se debe reducir la dimensión del problema lo más posible para generar una buena vista minable, la dificultad de hallar un modelo que represente un proceso aumenta mientras más variables se encuentren en el problema. Así que en este paso se reduce la cantidad de variables a sólo las necesarias para modelar el proceso en estudio.

- Realizar análisis estadísticos para reducir variables que posean una alta relación lineal, como por ejemplo un análisis de correlación.
- Identificar las posibles variables que se pueden reducir.
- Justificar la reducción de las mismas
- Construir la nueva vista minable con las nuevas variables reducidas

d) Actividades a realizar

Por parte de los investigadores:

Actividad: Realizar el proceso de limpieza, transformación y reducción de la vista minable con datos.

3. Fase 3: Desarrollo de herramientas de MD

Esta fase busca generar una herramienta de *software* que permita utilizar el modelo de MD, basándose en los requerimientos no funcionales. Las etapas de esta fase se muestran en la Figura 5.

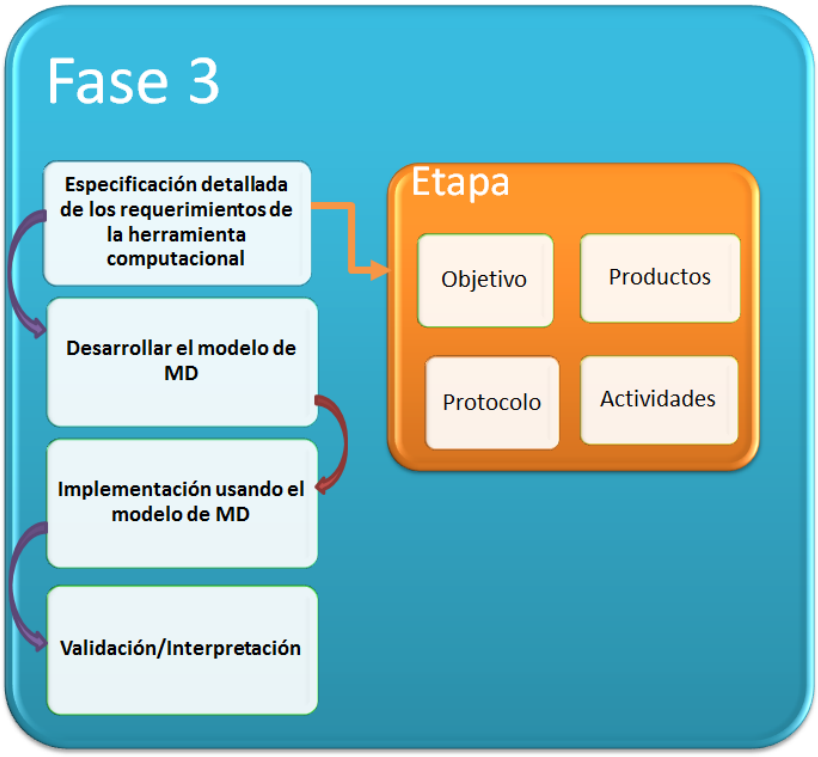

Figura 5: Etapas que conforman la fase 3.

3.1. Especificación detallada de los requerimientos de la herramienta computacional

a) Objetivo:

Esta etapa tiene como finalidad captar los requerimientos no funcionales, ya que los funcionales fueron descritos con los escenarios futuros deseados del punto 1.3.6.

b) Producto principal

Documento que contiene los requerimientos no funcionales mínimos para poner en funcionamiento la herramienta de MD.

c) Protocolo

En el paso 1.3.7 se hizo una especificación general de los requerimientos. La captura de los requerimientos tiene como objetivo principal la comprensión de lo que los clientes y los usuarios esperan que haga el sistema. En particular, los requerimientos funcionales fueron captados mediante la técnica de escenarios futuros en los pasos 1.3.6 y 1.3.7. También, los no funcionales fueron preliminarmente consideraros en el paso 1.3.7.

Existen en la literatura metodologías especializadas para levantar los requerimientos no funcionales, para el desarrollo de un proyecto de ingeniería de software.

Entre los requisitos no funcionales a definir se encuentran:

- Requisitos de interfaz de usuario, como por ejemplo: Estándar de GUI, Distribución de la pantalla, Restricciones de resolución, Estándares de botones, Estándares de mensajes de error, shortcuts, entre otros que intervengan en la interfaz del usuario.
- Interfaces de software, como: Conexiones entre el producto y software externo (identificado por nombre y versión), Identificar la información que comparten los componentes.
- Requerimientos de desempeño, entre los cuales se encuentran: los Tiempos de respuesta, el volumen o tiempo de utilización, el número de usuarios concurrentes, el número de operaciones concurrentes, entre otras restricciones de tiempo para sistemas de tiempo real..
- Adicionalmente se pueden mencionar: de portabilidad, costos, rendimiento, accesibilidad, entre otros.

Nosotros proponemos, más que escoger una metodología para esas tareas, que se seleccione una que se adapte a las necesidades del proyecto a desarrollar. A continuación se listan algunas de ellas:

- El desarrollo rápido de aplicaciones o RAD (acrónimo en inglés de *Rapid Application Development*)
- El método de desarrollo de sistemas dinámicos (en inglés *Dynamic Systems Development Method* o DSDM)
- Scrum como marco de trabajo basado en un proceso iterativo e incremental
- El Proceso Unificado Racional (RUP del inglés *Rational Unified Process*)
- La programación extrema o *eXtreme Programming* (XP)
- *Lean software development*
- *Agile Unified Process* (AUP) versión simple del *Rational Unified Process* (RUP)

Entre los aspectos a tomar en consideración al momento de desarrollar la captura de requerimientos no funcionales, pertinentes para nuestra metodología, son:

d) Actividades

Por parte de la institución/empresa:

Actividad: Proporcionar a los investigadores información sobre los requerimientos no funcionales deseados para la herramienta.

Por parte de los investigadores

Actividad:

- Seleccionar la metodología que permitirá la adquisición de requerimientos.
- Generar un documento con todos los requerimientos capturados de la institución/empresa

Trabajo conjunto:

Actividad: Reuniones virtuales para definir los requerimientos no funcionales

3.2. Desarrollar el modelo de MD

Analizar según el escenario en estudio, las técnicas de MD que se adaptan al mismo y a la vista minable (conceptual y operativo).

a) Objetivo:

Esta etapa tiene como finalidad escoger el modelo de MD resultante de la comparación de varias técnicas para una misma tarea.

b) Producto principal

Modelo de MD que representa los datos del histórico estudiado.}

c) Protocolo

- Selección del Software para realizar las tareas de MD
- Escoger la técnica de MD para la tarea identificada.

  Para la selección de la técnica, desarrollar una tabla de comparación entre las técnicas probadas, para conocer cual se adapta mejor a la estructura de los datos.
- Definir cuáles son los datos de entrenamiento y de prueba dispuestos en la vista minable, dependiendo de la técnica de MD a ser usada varían los porcentajes de la muestra para la prueba.
- Comenzar a realizar pruebas sobre la vista minable, para ir llenando la tabla comparativa de las técnicas de MD.
- Definir una estrategia para la validación de la técnica seleccionada, aplicarla y observar el rendimiento.
- Realizar las correcciones necesarias
- Repetir el procedimiento de ser necesario

d) Actividades

Por parte de los investigadores

Actividad: realizar los procesos necesarios para la escogencia del modelo de MD.

3.3. Implementación usando el modelo de MD

a) Objetivo:

Realizar la herramienta de MD con el modelo seleccionado.

b) Producto principal

Herramienta de MD.

c) Protocolo

Se desarrolla la herramienta computacional, cumpliendo con los requerimientos (no funcionales) adquirido en el punto 3.1 e integrando el modelo de MD generado en el punto 3.2. Este punto es prácticamente en su totalidad a ser realizado por parte de los investigadores para desarrollar la herramienta que se usará en tiempo real, de manera que cumpla con todas las especificaciones que se capturaron con los requerimientos no funcionales, para así pasar al siguiente punto de validación.

d) Actividades

Por parte de los investigadores

Actividad: realizar el desarrollo de la herramienta de MD.

3.4. Validación/Interpretación

a) Objetivo:

Validar la herramienta de MD.

b) Producto principal

Herramienta de MD validada.

c) Protocolo

En este etapa lo que desea es la validación de la herramienta con los expertos del sistema. A diferencia de la validación del algoritmo realizada en el paso 3.2, donde se verifica que el modelo generado cumpla con las expectativas, en este paso se trabaja directamente con los expertos para validar que la herramienta cumpla con las especificaciones de los requerimientos no funcionales, para ello se pueden plantear técnicas como: evaluaciones, inspecciones y tutoriales. De encontrarse algún error o mal funcionamiento se deben realizar las correcciones necesarias y volver al validar hasta que funciones de buena manera todas las especificaciones.

d) Actividades

Por parte de los investigadores

Actividad: realizar el proceso de validación de la herramienta de MD.

Por parte de la institución/empresa

Actividad: Realizar preguntas a los expertos para verificar que la herramienta cumpla con lo esperado por ellos.

Algunas preguntas que deben hacerse los expertos son:

- ¿Es esto lo que se especificó?
- ¿Cumple la herramienta con todas las especificaciones?
- ¿Cada especificación está funcionando correctamente en la herramienta?