MIDANO-AdD:

Metodología para la especificación de Tareas de Analítica de Datos

Aguilar, Jose
Universidad de los Andes
aguilar@ula.ve

Julio 2016.

Antecedente:

Actualmente, las organizaciones poseen grandes cantidades de datos que no son utilizados eficientemente. Desde dichos datos 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. MIDANO-AdD es una metodología que es una extensión a otra metodología, llamada MDANO, que sirve para identificar donde extraer conocimiento en una organización. Esta extensión es específica para identificar donde desarrollar Tareas de Analítica de Datos en una organización.

Objetivo

Desarrollo de herramientas computacionales de Analítica de Datos basadas en técnicas inteligentes o estadísticas (Redes Neuronales, Computación Evolutiva, Lógica Difusa, etc.), para la extracción de conocimiento desde bases de datos organizacionales

Metodología

Esta metodología está diseñada para el desarrollo de aplicaciones basadas en Analítica de Datos (AdD) para un proceso de cualquier institución/empresa. Está compuesta por tres fases:

(i) Identificación de fuentes para la extracción de conocimiento en una organización,
(ii) Preparación y tratamiento de los Datos y
(iii) Desarrollo de las tareas de AdD.

En la Figura 1 se observa el flujo de desarrollo de dichas fases.

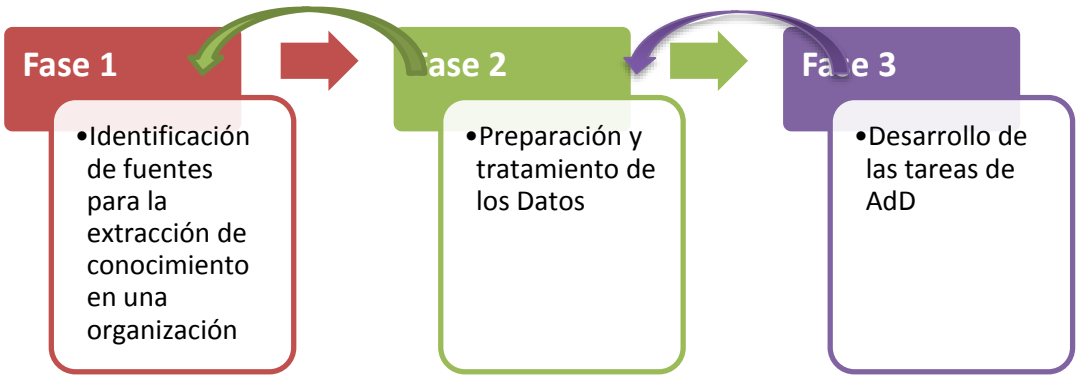

Figura 1: Flujo de desarrollo de la metodología propuesta.

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.

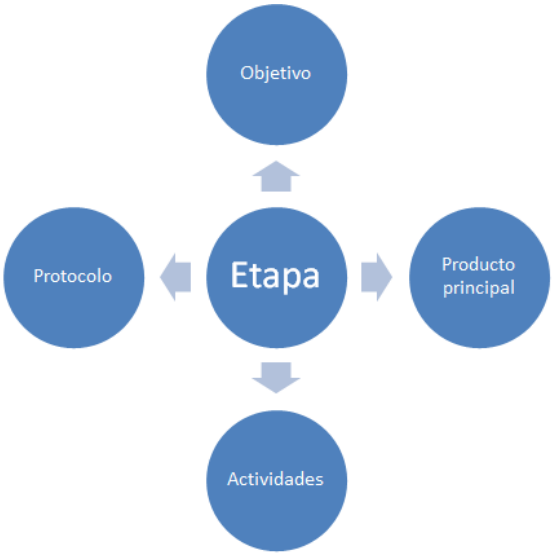

Figura 2: Aspectos que conforman cada etapa de las fases de la metodología.

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

a. *Objetivo*: describe cual es la meta que se quiere cumplir en la etapa respectiva
b. *Producto principal*: que es lo que se debe producir, en concreto, al final de la etapa.
c. *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.
d. *Actividades*: describe las tareas que se designan al grupo de ingeniería de conocimiento y a miembros de la organización/empresa, para lograr el objetivo de la etapa.

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 AdD.

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 AdD 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.

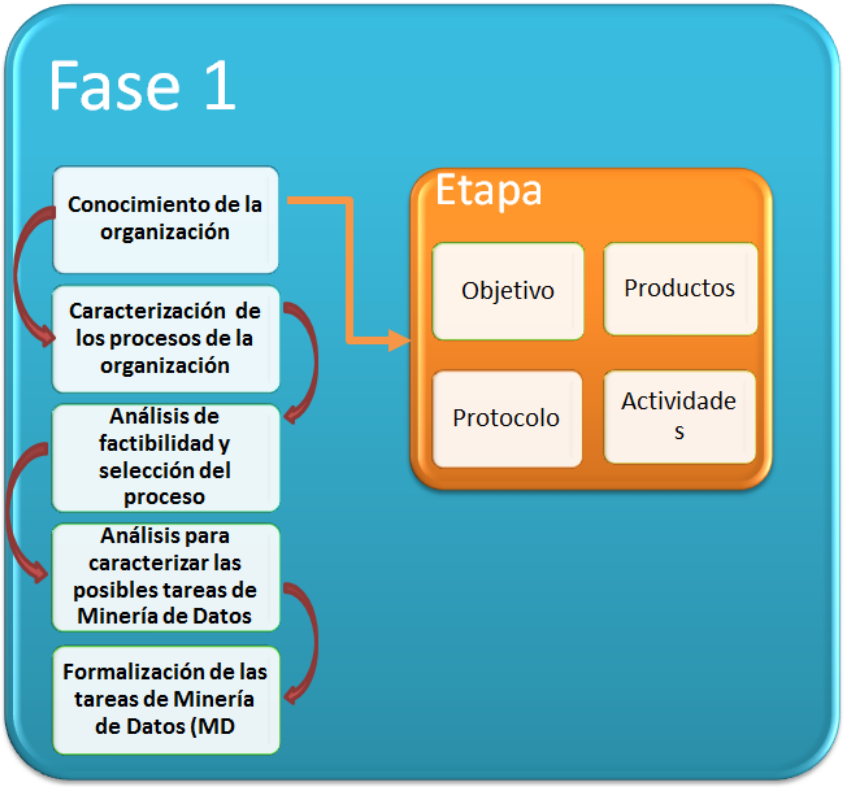

Figura 3: Etapas que conforman la fase 1.

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.

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 de la Tabla 1.

Tabla 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. En caso de tener un documento equivalente, identificarlo y colocarlo como los documentos bases (por ejemplo: documento organizacional, organigrama de la institución/empresa, etc.).
  Momento: Primera actividad que realiza la institución/empresa.
- Por parte de los ingenieros de conocimiento:
  Actividad: Estudio de la institución/empresa con la información proporcionada por la misma.
  Momento: Una vez consignado el documento por la institución/empresa.
- Trabajo conjunto:
  Actividad: Planificación de la primera entrevista, para que el grupo de ingenieros de conocimiento aclare dudas que tiene acerca de la organización, conozca mejor los procesos descritos en el documento, etc. El grupo de ingenieros de conocimiento podrá 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 entrevista.

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 AdD. 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 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. En caso de tener documentos equivalentes, facilitárselos al grupo de ingenieros de conocimiento.

  Momento: Previo a la primera visita.
- Por parte del grupo de 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 entrevista.

- Trabajo conjunto:
  Actividad: Entrevista para aclarar las dudas y preguntas, que generó el grupo de ingenieros de conocimiento.
  Momento: Durante la primera entrevista.

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 tareas de AdD sobre cada uno de ellos. Para ello, se utilizan criterios que permitirán la selección de uno o más procesos que cumplan con las características necesarias para la aplicación de tareas de AdD.

b) Producto principal:

Tabla con la evaluación de los procesos de la organización.

c) Protocolo:

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

1.3.1. Revisión de los procesos organizacionales
- Análisis detallado de los documentos proporcionados por la institución/empresa
- Determinación del propósito de aplicar AdD en los procesos
- Revisión de la literatura existente acerca de procesos semejantes, que se hallan tratado con AdD

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

1.3.3. Disponibilidad de grupos de expertos en la organización, en el proceso
- Verificar cuál es la disponibilidad de los expertos, y sus intereses en el proceso.

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 AdD.
- 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 AdD

- 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 AdD, se usan criterios como los descritos en la Tabla 2.

Tabla 2. 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 AdD	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.
Qué información se recoge del proceso para ser almacenada	Cantidad de información que recoge el proceso.
Con qué 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

Con qué 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 2 de cada proceso de la institución/empresa, y se realiza la 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 dé como 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 AdD.

d) Actividades:

- Por parte de los ingenieros de conocimiento:
  Actividades: Estudio de los procesos con la información proporcionada.
  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 la segunda visita.

1.4. Caracterizar las posibles tareas de Analítica de Datos

a) Objetivo:

Una vez conocida la empresa/institución, recabada información de la misma, analizados sus procesos, y seleccionado el (los) proceso(s) sobre el(los) cual(es) se realizará la tarea AdD por su importancia para la empresa/institución, en esta etapa se procede a caracterizar las posibles tareas de AdD a realizar en los procesos organizacionales priorizados (objetivos, requerimientos, factibilidad, etc.), con la finalidad de escoger las tareas de AdD 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 AdD, tablas de cada tarea de AdD.

c) Protocolo

Los ingenieros de conocimiento realizarán una serie de preguntas a personas adecuadas en la institución/empresa, para obtener las necesidades de la institución, acorde al proceso escogido previamente. Para caracterizar las posibles tareas de AdD 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 los cuales se da una descripción general de los resultados esperados o deseados, que se pueden obtener después de aplicar la tarea de AdD al escenario actual. Así, en esta etapa se realizan los siguientes pasos:

1.4.1 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?

1.4.2 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. Las preguntas a los expertos para caracterizar las variables de un proceso son:

- ¿Cuál es el flujo de actividades detallado del proceso en estudio?.

- Enriquecer el diagrama de actividades, usando 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 en 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 3, 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 3. 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 AdD 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 AdD a aplicar):

Tabla 4. 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	Actor(es) que interviene(n) para el	Variables que están relacionadas	Actividades que se realizan para	Funcionalidades que no tiene el actor(es)/actividades/proceso en este momento, pero que pudieran tener

obtener por medio de AdD	desarrollo del producto	con el producto	obtener el producto	

- Selección de los escenarios factibles de AdD. 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). 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, para la elaboración de 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 pueden usar criterios como los señalados en la Tabla 5.

Tabla 5.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

1.4.3 Especificación de los requerimientos para el plan tecnológico de desarrollo del(los) escenario(s) futuro(s) (tarea(s) de AdD 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 AdD en el escenario escogido?
  - ¿Qué interacción tendrá la tarea de AdD de datos con los actores del escenario?
  - ¿Qué interacción tendrá la tarea de AdD de datos con el escenario actual del proceso escogido?
  - ¿Qué interacción tendrá la tarea de AdD 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 AdD en la plataforma?
  - Identificar los datos de entrada para la tarea de AdD, y las herramientas/sistemas con que cuenta la organización para proveerlos (pueden ser datos de entrada o salida del proceso)

Tabla 6. 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 asociados:	De existir requerimientos relacionados con el mismo, colocar el(los) Id(s) del(los) requerimiento(s).

1.4.4 Elaboración de los casos de uso para los requerimientos funcionales

- Generar los casos de uso que solventarán los requerimientos funcionales especificados. Dichos casos de uso serán diagramados usando UML.

1.4.5 Elaborar el plan preliminar de actividades para el desarrollo de herramienta de AdD.

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 AdD.

  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 Analítica de Datos

a) Objetivo

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

b) Producto principal

Documento formal con la definición del problema.

c) Protocolo de etapa

Desarrollo de un informe por parte del grupo de *ingenieros de conocimiento*, con la conceptualización de los procesos a estudiar, y la caracterización de sus problemáticas operacionales y del uso de la AdD en dichos procesos. Este documento es la definición inicial de la tarea de AdD, que irá evolucionando a medida que el proyecto avance. El documento contendra las tablas con las especificaciones de cada tarea de AdD, basada en la tabla 7.

Tabla 7. Tabla para describir tareas de AdD.

Nombre de la tarea	<nombre de la tarea>
Descripción	<La finalidad de esta tarea>
Fuente de datos	<BD, historicos>
Tipo de tarea de analítica de datos	<Asociacion, Agrupamiento, Clasificacion, Predicción, reglas de asociación, etc.>
Técnicas de analítica de datos	<Define las posibles tecnicas a usar, por ejemplo: regresión, redes neuronales artificiales, algoritmo K-NN, etc.>
Tipo de Modelo de Conocimiento	<modelo descriptivo, modelo prescriptivo, modelo de optimizacion, modelo predictivo, etc.>
Tareas relacionadas de analítica de datos	<Con que otras tareas de AdD se relaciona>
Tipo de tarea del ciclo autonómico	<Pueden ser para observar, analizar/interpretar, o actuar sobre el proceso>

d) Actividades

- Por parte de los ingenieros de conocimiento:
  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 AdD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema de estudio. Esto conlleva a realizar distintas operaciones con los datos, con la finalidad de acondicionarlos para desarrollar un modelo de AdD. Para realizar este proceso se crean diferentes vista minables, que básicamente contienen la información de las variables y los datos del historial a ser usados por la tarea de AdD. A continuación se definen algunos conceptos usados en esta fase:

- *Una vista minable conceptual* describe en detalle cada una de las variables a ser tomadas en cuenta por la tarea de AdD. 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).
- *Una vista minable operativa* es el almacén de datos construido a partir de la vista minable conceptual, en donde se cargan los datos desde las BD operacionales de la organización. A ese almacén también se le denominado el *Data Mart* (o *Data Warehouse*, según el tamaño del almacén) de la tarea de AdD, y normalmente es un modelo de datos multidimensional del tipo estrella.

Esta fase realiza la preparación y tratamiento adecuado de los datos, que conforman la *vista minable operativa*, que serán utilizados para el desarrollo de la tarea de AdD. En la Figura 4 se muestran las etapas que conforman esta fase.

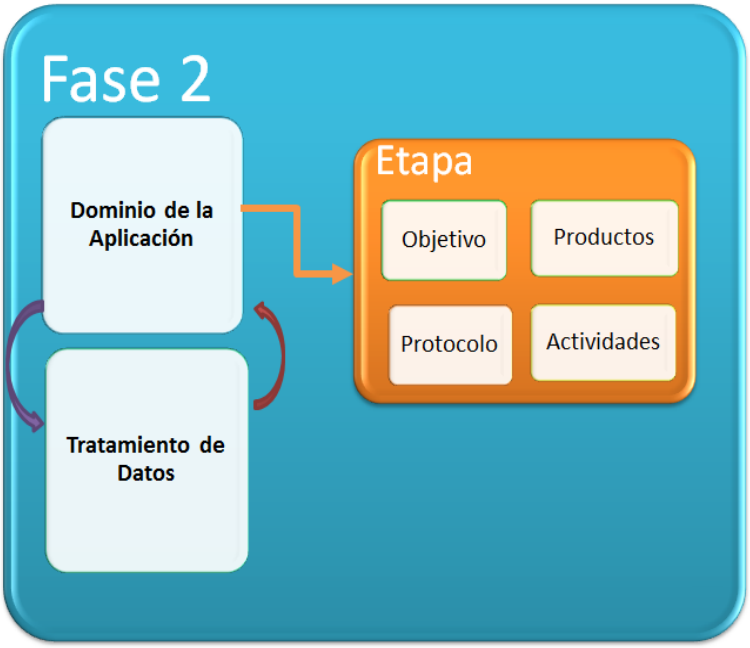

Figura 4: Etapas que conforman la fase 2.

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 ((modelo multidimensional), de interés para la tarea de AdD.

Otros objetivos serian:

- Ubicar y comprender los datos asociados a las tareas de AdD
- Construcción de la tabla con las operaciones de (E)xtracción, (T)ransformación y Carga (L), para las variables identificadas en la vista minable conceptual
- 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
- Tabla ETL
- Vista minable operativa (modelo multidimensional)
- Descripción de la(s) variable(s) objetivo(s)

c) Protocolo de etapa

2.1.1. Comprensión de los datos de entrada

Según la tarea de AdD 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 a la tarea de AdD, 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 la tarea de AdD 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 fórmulas, entre otras variables), se establecen los campos adicionales, etc. Puede suceder que otras tareas de AdD tengan esas variables, aquí se identifican, para integrarlas en una única vista minable de todas las tareas de AdD a desarrollar.
- 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.

La tabla 8 muestra un ejemplo de la información que contendrá la vista minable conceptual.

Tabla 8. Vista Minable Conceptual

Variable	Descripción	Procedencia	Observaciones

2.1.3. Definir las variables objetivos

Una vez planteado el escenario futuro y la tarea de AdD a realizar, es preciso detectar las variables que permitirán la consecución de los objetivos de AdD. 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 AdD. Así, en esta fase se desea definir las variables objetivo, y ubicar dichas variables en la vista minable descrita previamente. 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)

2.1.4. Integración de los datos de entrada

Una vez obtenida la vista minable conceptual, se procede a diseñar el modelo de datos multidimensional, para poder cargar los datos de las bases de datos operacionales, convirtiéndose en la vista minable operativa. El modelo multidimensional está compuesta por dos tipos de tablas:

- *Tablas de Hecho*: contiene la información/modelo de conocimiento generada por la tarea de AdD, además de las claves para accesar a las diferentes tablas de dimensiones.
- *Tablas de Dimensiones*: Define los datos organizados lógicamente por temas específicos, a los cuales se acceden vía claves.

Las siguientes tablas se deben definir, para especificar la vista minable operacional, para cada tabla de dimensión y de hecho requerida por la tarea de AdD.

Tabla 9. Tabla de Hecho

Nombre	Nombre de la tabla de hecho
Variables Objetivos	Variables que describen o se asocian al conocimiento extraido (predicciones, etc.)
Claves a las tablas de dimensiones	Todas las claves a las tablas de dimensiones
Otras variables	Variables requeridas por la tarea de AdD, por ejemplo, derivadas de operaciones de procesamiento de las dimensiones

Tabla 10. Tabla de Dimensión

Nombre	Nombre de la tabla
Claves de la dimensión	Claves a la tabla de dimensión
Otras variables	Variables que describen el tema asociado a esa dimensión

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: identificar datos comunes, determinar tipo de fusión (enlazado, unión, etc.), qué tipo de dato va a quedar, qué nombres van a quedar, formato de integración, etc.
Así, todos los datos que las tareas de AdD manejarán deben estar en un mismo repositorio (Data Mart). La integración de estos datos debe darse en un repositorio (físico o digital), a la 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).

Para completar la vista minable operativa, se deben especificar todas las operaciones de (E)xtracción, (T)ransformación y Carga (L), para cada una de las variables de la vista minable conceptual. Para ello, se llena la Tabla 11, que específica para cada tarea de que fuente de datos organizacional se extraerá, cuales tipos de transformaciones se les realizará (limpieza, estudios de dependencia, etc.), y para que dimensión del modelo multidimensional (vista minable operativa) irá.

Tabla 11. Tabla ETL

Variable	Extracción	Transformación	Carga
	De que fuente de datos organizacional se extraera	Aqjui se indican todo el proceso de pre-procesamiento de los datos (estudios de dependencia, limpieza, cambio de formatos, etc.)	A que dimensión del modelo de datos irá

d) Actividades

- Por parte de la institución/empresa:
  - ➢ Proporcionar la información de los datos asociado a la tarea de AdD
  - ➢ Proveer los datos asociados a la vista minable conceptual provenientes de los servidores de la institución/ empresa

- Por parte de los ingenieros de conocimiento:
  - Generar una descripción de los datos y las relaciones que tienen con las variables.
  - Conocer como están almacenados los datos
  - Construir la vista minable conceptual
  - Construir la vista minable operativa
  - Construir la tabla ETL
  - Seleccionar y ubicar la(s) variable(s) objetivo(s) en la vista minable con datos
- Trabajo conjunto:
  Reuniones virtuales para completar/validar esta etapa.

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. Esas operaciones son descritas en la tabla ETL, en la etapa anterior.

b) Productos principales

- Vista minable operacional (Data Mart) lista para usar

c) Protocolo de etapa

La vista minable operativa es preparada mediante herramientas especializadas en 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 según 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 se 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 puede 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 la tarea de AdD.

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). Para ello se debe:

- 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 tarea de AdD, y limitaciones técnicas tales como los volúmenes máximos de datos permitidos. Se debe reducir la dimensión lo más posible, para generar una buena vista minable. La dificultad de una tarea de AdD puede aumentar mientras más variables innecesarias se usen. Así, en este paso se reduce la cantidad de variables, a sólo las necesarias para modelar el proceso en estudio.

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

d) Actividades a realizar

Por parte del grupo de ingeniería de conocimiento:

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

3. Fase 3: Desarrollo de la tarea de AdD

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

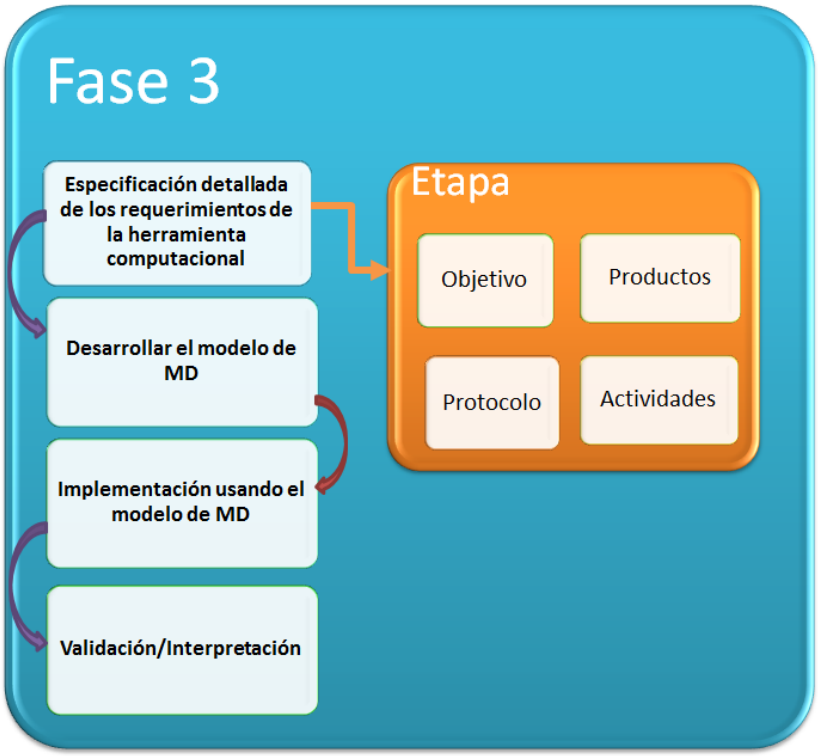

Figura 5: Etapas que conforman la fase 3.

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 AdD.

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. En general, 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.

d) Actividades

- Por parte de la institución/empresa:
  Actividad: Proporcionar los requerimientos no funcionales deseados para la herramienta.

- Por parte grupo de ingeniería de conocimiento:
  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 Conocimiento de la Tarea de AdD

Analizar según el escenario en estudio, lo especificado en la Tabla 7, y las vistas minables (conceptual y operativo), las técnicas de AdD que se adaptan mejor. A partir de allí, desarrollar el modelo de conocimiento definido en la tarea de Analítica de Datos

a) Objetivo:
Esta etapa tiene como finalidad, desarrollar el modelo de conocimiento de la tarea de AdD.

b) Producto principal
Modelo de conocimiento (de optimización, de identificación, etc.).

c) Protocolo

- Selección del Software para realizar las tareas de AdD
- Escoger las técnicas de AdD 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 operacional. Dependiendo de la técnica de AdD a ser usada, varían los porcentajes de la muestra para la prueba.
- Comenzar a realizar pruebas, para ir llenando la tabla comparativa de las técnicas de AdD.
- 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 del grupo de ingeniería de conocimiento
Actividad: realizar los procesos necesarios para la escogencia del modelo de AdD.

3.3. Implementación de la herramienta de toma de decisiones usando el modelo de AdD

a) Objetivo:

Realizar la herramienta de toma de decisiones usando el modelo de conocimiento generado por la tarea de AdD con.

b) Producto principal

Herramienta de toma de decisiones.

c) Protocolo

Se desarrolla la herramienta computacional, cumpliendo con los requerimientos (no funcionales) adquirido en el punto 3.1 e integrando el modelo de AdD generado en el punto 3.2. Este punto es realizado por parte del grupo de ingeniería de conocimiento, y debe cumplir con todas las especificaciones que se capturaron con los requerimientos no funcionales, para así pasar al siguiente punto de validación.

d) Actividades

Por parte del grupo de ingeniería de conocimiento
Actividad: realizar el desarrollo de la herramienta de toma de decisiones.

3.4. Validación/Interpretación

a) Objetivo:

Validar la herramienta de toma de decisiones.

b) Producto principal

Herramienta de toma de decisiones validada.

c) Protocolo

En este etapa se valida 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 a validar hasta que funcione de buena manera.

d) Actividades

Por parte del grupo de ingeniería de conocimiento

Actividad: realizar el proceso de validación de la herramienta.

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?