CMIN 1.0 – Herramienta CASE para el soporte de proyectos de Minería de Datos basado en CRISP-DM

Jhon Emmanuel Zuñiga Paredes
Juan Carlos Guarin Walteros

Director: MSc. Carlos Alberto Cobos Lozada

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Grupo de I+D en Tecnologías de la Información
Gestión de la Información – Minería de Datos
Popayán, Abril de 2009

Agradecimientos

A la Universidad del Cauca, al Programa de Ingeniería de Sistemas y a los docentes que durante toda la carrera de Ingeniería de sistemas desarrollaron en nosotros cualidades como la recursividad en busaca de soluciones, la mejora continua y sobre todo la Ingeniería como profesión y método de vida.

Al MSc. Carlos Alberto Cobos Lozada por tenernos en cuenta para desarrollar este proyecto, dado que este trabajo de grado es una de las tantas iniciativas que lo han destacado tanto en el ambiente local como internacional. Además por sus consejos y enseñanzas en el transcurso del Proyecto como docente, ingeniero y amigo.

A nuestras familias por su total apoyo y cariño necesario para derribar los obstáculos presentados en el proyecto.

Y finalmente a todas las personas que de una u otra forma participaron en el desarrollo de este proyecto.

Mil Gracias a todos.

Dedicatoria

Especialmente a mi padre que aunque ahora esta descansando en el cielo, me enseño a luchar por mis sueños sin importar lo difícil que sea llegar a ellos.
A mi madre que me enseño que con amor y consistencia se pueden lograr nuestras metas.
A mis hermanos y sobrinos que me brindaron su apoyo y compañía en momentos difíciles.
A mi esposa que llego a mi vida a llenarla de amor e ilusión, brindándome el último empujón para lograr esta gran meta.
A mi hija que llego y me cambio la vida total y hermosamente, lo cual me impulso a ser cada día mejor persona para poder ser un excelente padre como lo fueron mis padres.
Y finalmente a Dios por brindarme esta hermosa familia que estuvo a mi lado y me apoyo siempre.

Juan G.

Tabla de contenido

Lista de Figuras

Lista de Tablas

PARTE 1: CONTEXTO DE LA INVESTIGACIÓN

1 PLANTEAMIENTO DEL PROBLEMA

1.1 Definición

La minería de datos es una tecnología que se utiliza principalmente para obtener información previamente desconocida en grandes volúmenes de datos, y en los últimos años se han reportado excelentes resultados en su uso en diversas áreas de la ciencia y la tecnología, como por ejemplo en medicina [1], mercadeo [2] y agricultura [3].

En este sentido, Colombia no es la excepción. Actualmente, existen diferentes iniciativas que utilizan la minería de datos en las organizaciones y otras que esperan convertirla en un tema central de investigación, ejemplos de esto son las investigaciones realizadas en la Universidad del Valle [4][5][6], la Universidad Nacional de Colombia sede Manizales [7] y la Universidad de Antioquia [8], los convenios entre empresas y universidades para desarrollar proyectos en este tema [9] y las estrategias que esta aplicando actualmente el gobierno en el área de tecnología[10].

Lo anterior es un indicio serio de que en nuestro país se ha detectado la necesidad de contemplar esta tecnología y explotar sus diferentes ventajas. Pero, para lograr un verdadero desarrollo en este tema hace falta realizar diferentes estrategias, entre ellas, el desarrollo de jornadas masivas de difusión alrededor del tema y proyectos relacionados y tratar formalmente los problemas que enfrentan los desarrolladores de proyectos de esta tecnología, en especial las "Herramientas de desarrollo". Con respecto a las herramientas, tema central de este proyecto, se han determinado varios inconvenientes [11], entre ellos:

- La integración de diferentes técnicas de minería, algunas solo cuentan con un par de técnicas.
- Extensibilidad, las herramientas no permiten la adición dinámica o flexible de técnicas de minería de datos, para aumentar su poder de solución.
- Integración con estructuras de almacenamiento de datos, no realizan el procesamiento sobre las bases de datos para procesamiento transaccional, si no que es necesario la importación continua de los datos.
- Soporte para usuarios expertos tanto como para usuarios novatos.
- Administración de cambios de los datos, es decir reconocer cuando un modelo ya no es valido para un problema específico debido al cambio de comportamiento de los datos.

Enfocados en el contexto regional, en este proyecto se valoraron otros aspectos de las herramientas más reconocidas en la actualidad, y que sirven de complemento para establecer una lista de características deseables en una herramienta de desarrollo que soporta un proyecto de minería de datos, ellas son:

- Acceso: Ciertas herramientas son de difícil acceso por el alto costo de adquisición, el cual es difícil de pagar para la gran mayoría de empresas Colombianas, que se caracterizan por ser micro, pequeñas y medianas empresas (MIPYMES). Aunque existen casos específicos en los que se han utilizado [12].

- Interfaz amigable: No cuentan con un método interactivo para guiar al usuario en el proceso de minería de datos, lo que es una dificultad para usuarios que tienen poco conocimiento del proceso.

- Metodología: No se basan en una metodología estándar para seguir el proceso de minería de datos, sino por el contrario usan una propia, lo que es un inconveniente, debido a que actualmente existen metodologías estándar para el proceso de minería de datos que pretenden facilitar la realización de nuevos proyectos con características similares, optimizar la planificación y dirección de los mismos, reducir su complejidad y permitir realizar un mejor seguimiento a estos [13], entre las que se destacan CRISP-DM (Cross – Industry Standard Process for Data Minning) [14] y SEMMA (Simple, Explore, Modify, Model, Assess) [15]. SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que CRISP-DM, mantiene como foco los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase donde SEMMA comienza realizando un muestreo de datos, mientras que CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista se puede considerar que CRISP-DM está más cercana al concepto real de un proyecto de minería de datos, pudiendo ser integrada con una Metodología de Gestión de proyectos específica que completaría las tareas administrativas y técnicas, además CRISP-DM es de libre distribución sin ningún costo a diferencia de SEMMA[13].

- Extensibilidad: No cuentan con la capacidad de adicionar dinámicamente algoritmos al conjunto de técnicas que se entregan en cada versión. Las herramientas evolucionan con versiones, al ritmo de las empresas que las producen y no propician un desarrollo en comunidad similar al de software libre, excepto por Weka (en una forma un poco compleja para el usuario de la herramienta).

- Equipos: El trabajo en equipos de desarrollo en proyectos de minería de datos, se puede ver dificultado debido a que la gran mayoría de las herramientas existentes soportan flujos de trabajo para un solo usuario por proyecto, siendo esto un problema para proyectos de gran envergadura.

Del listado de las herramientas más representativas, según MetaGroup [16] y kdnuggest [17], se valoraron Clementine[18], Insightful Miner[19], WEKA [20], CART [21], PolyAnalyst [22] y SAS [23] (ver Tabla 1. Valoración de herramientas de minería de datos.). Para ver más detalles del estudio comparativo, refiérase a [16], [17] y muy especialmente a [24].

Herramienta	Acceso	Desarrollo Interfaz	Metodología	Extensibilidad	Equipos
Clementine	Costoso	Sencillo	Parte de CRISP-DM	No, sólo con el nuevo release	No
Insightful Miner	Costoso	Regular	PROPIA	No, sólo con el nuevo release	No
WEKA	Libre	Complejo	PROPIA	Si, por un	No

				usuario experto	
CART	Costoso	Regular	PROPIA	No, sólo con el nuevo release	No
PolyAnalyst	Costoso	Regular	PROPIA	No, sólo con el nuevo release	No
SAS	Costoso	Sencillo	SEMMA	No, sólo con el nuevo release	No

Tabla 1. Valoración de herramientas de minería de datos.

Además y teniendo en cuenta que las herramientas CASE (Computer Aided Software Engineering) son usadas para estandarizar procesos de desarrollo y automatizar la mayoría de metodologías existentes para los procesos de desarrollo de software [25] y que en especial las CASE integradas (WorkBench Integrated) apoyan las actividades del proceso de software, a través de una interfaz consistente y homogénea, y una administración centralizada de la documentación [25]; se plantea que una estrategia para que la región adopte y use la minería de datos en pro de la mejora de sus procesos, podría ser la creación de una herramienta CASE integrada de minería de datos que se base en una metodología estándar como CRISP-DM (teniendo en cuenta las diferencias vistas anteriormente entre las dos metodologías mas representativas, SEMMA y CRISP-DM), y que para su modelado y construcción contemple las mejores características de las herramientas existentes y supere algunos de sus inconvenientes.

Por lo anterior en este proyecto se busco dar solución a algunos de los problemas detectados en las herramientas de minería de datos existentes, planteando como pregunta de investigación la siguiente: ¿Cómo debe ser modelada y construida una herramienta CASE integrada basada en una metodología estándar, que presente una guía interactiva y útil al usuario (ofreciendo ayuda y soporte en cada paso del proyecto), que cuente con la posibilidad de adicionar dinámicamente técnicas de minería de datos (flexible y extensible) y sea asequible para empresas de la región y el mundo (de libre distribución)?.

1.2 Justificación

En nuestra región la investigación y utilización de la minería de datos es aún muy baja, debido en parte, a la poca formación que se tiene en el tema en las instituciones de educación superior que ofrecen programas de ingeniería de sistemas o afines a nivel profesional o tecnológico, como lo son entre otras, la Universidad Cooperativa de Colombia [26], Colegio Mayor del Cauca [27] y la Fundación Universitaria de Popayán [28] las cuales no tienen en sus planes de estudio la temática de minería de datos. Por otro lado, recientemente en la Universidad del Cauca, se han estado desarrollando trabajos de grado alrededor de esta temática (reglas de asociación para comercio electrónico, minería de datos en tiempo real, clustering para recuperación de información web, minería de datos web y otras técnicas de minería aplicadas al sector agroindustrial) y en el segundo semestre del 2007 se abre por primera vez un curso electivo.

El modelado y desarrollo de una herramienta CASE integrada (nueva tecnología), de fácil acceso y orientada al usuario final, apoyará a diversas organizaciones y personas

interesadas en esta tecnología, gracias a que podrán utilizar una herramienta a su alcance, la cual les facilitará el desarrollo de proyectos de esta índole, permitiéndoles poner en práctica sus conocimientos así como investigar diferentes maneras de cómo beneficiar nuestra región con esta tecnología.

La investigación necesaria para desarrollar este proyecto permitirá apropiar conocimientos en el área de minería de datos, brindando información valiosa a diferentes grupos que están comenzando a desarrollar proyectos e investigaciones relacionadas con este tema, permitiéndoles obtener una mejor calidad tanto en proyectos actuales como futuros.

El proyecto ha establecido la mejor forma de desarrollar (proceso y arquitectura) una herramienta tecnológica, que permite al usuario final desarrollar proyectos de minería de datos, guiado de una manera dinámica en una de las metodologías más aceptadas y reconocidas internacionalmente en la minería de datos como lo es CRISP-DM, la cual le permitirá el desarrollo de proyectos de una manera más organizada y coherente, además de aprender esta metodología mientras trabaja en ella, permitiéndole lograr obtener resultados más confiables en cada proyecto realizado.

Por último, al brindar a la comunidad local, regional y nacional una herramienta económicamente accesible por las diferentes organizaciones, se espera motivar el desarrollo de más proyectos de minería de datos con un nivel bajo de riesgo. Esta herramienta podrá ser usada en la Especialización en Desarrollo de Soluciones Informáticas de la Universidad del Cauca, en el curso de Minería de Datos, al igual que en la Maestría en Ciencias de la Computación y en el pregrado en Ingeniería de Sistemas, logrando con ello, iniciar un proceso de masificación en el uso de la herramienta.

1.3 Antecedentes

En la sección 1.1 relacionada con la definición del problema ya se señalaron algunos inconvenientes de las herramientas de minería de datos más usadas en la actualidad, además en la misma sección se presentó como este proyecto pretende dar solución a algunos de esos inconvenientes, destacando el aporte e innovación del proyecto.

Como otro antecedente importante, es preciso citar el caso de la empresa "DaimlerChrysler's Data Mining" [29], que participa en el proyecto CRISP-DM. Esta empresa creó una instancia de CRISP-DM para soluciones de marketing llamada "Quick Reference Guide". "Quick Reference Guide" fue integrada con una herramienta software que permite la documentación y recolección de lecciones aprendidas en cada etapa. Además utilizaron un módulo de Clementine para la etapa de modelado, donde establecen plantillas de los modelos que pueden ser solución para diferentes problemas de marketing. La experiencia que destaca "DaimlerChrysler's Data Mining" se centra en que la utilización de la instancia de CRISP-DM facilita el inicio de un proyecto, optimizando el resultado de cada uno de estos con las lecciones aprendidas y las plantillas.
Con relación a este trabajo y teniendo en cuenta las ventajas que puede ofrecer realizar una instancia de la metodología base CRISP-DM, CMIN ofrece la opción de

seleccionar las actividades que se desean realizar de cada fase, lo cual permite a los usuarios crear una instancia y si lo desea guardarla como plantilla para futuros proyectos.

2 CONTRIBUCIÓN A LA SOLUCIÓN

2.1 OBJETIVOS

A continuación se presentan los objetivos planteados para el proyecto, de la misma forma como aparecen en el documento del anteproyecto aprobado por el Comité de Investigaciones de la Facultad de Ingeniería Electrónica y Telecomunicaciones.

2.1.1 Objetivo General

Modelar, desarrollar y evaluar una herramienta CASE integrada para soportar y orientar el desarrollo de proyectos de minería de datos, cumpliendo totalmente con los pasos y tareas definidos en CRISP-DM y con la capacidad de extender su funcionalidad dinámicamente.

2.1.2 Objetivos Específicos

1. Modelar y Desarrollar una herramienta CASE integrada que le permita al desarrollador de un proyecto de minería de datos:

    - Realizar el seguimiento de su proyecto en cada una de las fases definidas en su instanciación del proceso CRISP-DM.
    - Desarrollar de una forma fácil y práctica la labor de modelado, selección y prueba de las técnicas de descubrimiento de conocimiento que aplican en un proyecto especifico de minería de datos.
    - Aprender y ampliar su conocimiento en cuanto al proceso de descubrimiento de conocimiento con Minería de Datos, a través de información detallada y apropiada de cada una de las etapas que se desarrollan en CRISP-DM.
    - Ampliar dinámicamente el conjunto de técnicas de minería de datos que ofrece la herramienta.

2. Desarrollar la herramienta CASE basado en una arquitectura multinivel y servicios web XML, usando los patrones de software que se ajusten a la misma en las diferentes etapas de desarrollo del proyecto y tomando como base UP como metodología de desarrollo y UML como notación estándar para la documentación.

3. Evaluar el grado de aceptación de la herramienta CASE a través de una experiencia piloto, basado en una prueba beta con estudiantes de la Universidad del Cauca.

2.2 RESULTADOS OBTENIDOS

Como resultado del proyecto se logró cumplir a cabalidad con los objetivos (obteniendo resultados adicionales a los originalmente propuestos). Los productos obtenidos son los siguientes:

- Herramienta CASE CMIN 1.0 la cual permite la definición de metodologías, para poder realizar proyectos de minería de datos basados en estas, lo cual permite obtener resultados mas organizados y exitosos. CMIN 1.0 contiene los siguientes módulos:

  - Modulo de procesos: el cual permite agregar los diferentes pasos de una metodología de manera jerárquica, para ser utilizada como guía en el desarrollo de proyectos de minería en CMIN 1.0.
  - Modulo de plantillas: el cual permite especializar una metodología creando plantillas de proyectos para un área específica de desarrollo humano (marketing, salud, ecuación, entre otras).
  - Modulo de proyectos: el cual permite el desarrollo y seguimiento de proyectos de minería de datos basándose en una metodología previamente definida.
  - Modulo de agregación dinámica de librerías: el cual permite agregar librerías de enlace dinámico (DLL) con la implementación de diferentes objetos necesarios para poder realizar tareas de minería de datos en el Work Flow ofrecido por CMIN 1.0.
  - Modulo de WorkFlow: el cual permite la aplicación de las técnicas de minería de datos en cada paso de la metodología de una manera más ordenada.
  - Modulo de reportes: el cual permite la definición de los reportes que se requieren para cada momento de ejecución de un proceso.

- Servidor que permite alojar nuevas implementaciones de algoritmos y metodologías de minería de datos, los cuales permiten a los usuarios actualizar CMIN 1.0 cuando sea necesario.

- Monografía de trabajo de grado: El presente documento, donde se describe el proceso que fue seguido para el desarrollo de la aplicación, las pruebas y sus resultados, los problemas, las respectivas soluciones, los lineamientos recomendados, los aportes fundamentales del proyecto, las conclusiones y el trabajo futuro que se puede seguir desarrollando.

- Presentación en evento internacional: En el marco del Microsoft Research Academic Summit realizado en Panamá del 14 al 16 de mayo de 2008, se realizó la presentación de un primer prototipo operacional de CMIN, obteniendo excelentes comentarios de los investigadores asistentes al evento y además se logró ser parte de los cinco proyectos seleccionados por Microsoft para salir en una noticia del canal CNN en el programa Adelantos (disponible en http://www.unicauca.edu.co/~ccobos/cnn-adelantos.wmv)

- Artículo que resume el contenido de la presente monografía que será presentado en el IEEE International Conference on Data Mining series (ICDM 2009) a realizarse en Miami, Florida del 6-9 de diciembre de 2009 y para el cual se debe entregar la versión definitiva en ingles antes del 26 de Junio de los corrientes.

PARTE 2: CONTEXTO TEÓRICO

3 Herramientas CASE

En la medida que las computadoras se utilizaban en diversas disciplinas para facilitar el trabajo, la demanda de software ha aumento dramáticamente. Para cubrir esta demanda, se crearon metodologías que buscan establecer un estándar de desarrollo, además se creo un soporte automatizado para estas metodologías, el cual se denomino "ingeniería del software asistida por computador" (CASE por sus siglas en Ingles)[30].

Las herramientas CASE ayudan a reducir el tiempo empleado en el desarrollo de un sistema, lo que mantiene el costo estable y contribuye en su calidad. [25]. Además, permiten al analista documentar y modelar un sistema, desde la definición de requerimientos hasta el diseño, implementación y prueba[25]. Diferentes investigaciones y usuarios sustentan que:[25]:

- Los usuarios de las CASE pueden mejorar la eficiencia, usando mecanismos de corrección de errores en el diseño de un sistema, usando además herramientas para crear los diagramas correspondientes y evitando con esto empezar desde cero para realizarlos nuevamente.

- Las organizaciones se motivan por la adopción de las CASE debido a que estas herramientas se basan en una metodología estándar para el desarrollo de sistemas, cuentan con un repositorio de datos central que describe el diseño del sistema reforzando el uso de la metodología soportada y ayudan a mejorar la calidad y seguridad durante el proceso de ingeniería de sistemas.

Componentes de las CASE: A continuación se describen los principales componentes de las CASE [30]:

- **Repositorio:** Base de datos central de una herramienta CASE. El repositorio incluye la información que se va generando a lo largo del ciclo de vida del sistema, como por ejemplo: componentes de análisis y diseño, estructuras de programas, algoritmos, etc. Las características más importantes de un repositorio son[30]: Tipo de información: metodología usada, datos, gráficos, procesos, informes, modelos o reglas. Tipo de controles: módulo de gestión de cambios, de mantenimiento de versiones, de acceso por clave, de redundancia de la información.
- **Módulos de diagramación y modelamiento:** Este modulo ofrece a la herramienta CASE la creación de los diversos tipos de diagramas y modelos necesarios en el apoyo de la metodología que esta siguiendo [30]. Algunos de los diagramas y modelos utilizados con mayor frecuencia son: Diagrama de flujo de datos, modelo entidad – interrelación, historia de la vida de las entidades, diagrama estructura de datos, diagrama estructura de cuadros y técnicas matriciales.
- **Herramienta de prototipado:** El objetivo principal de este componente es poder mostrar al usuario, desde los momentos iniciales del diseño, el aspecto que tendrá la aplicación una vez desarrollada. Ello facilitará la aplicación de los cambios que se

consideren necesarios, todavía en la fase de diseño, proporcionan una realimentación inmediata, que ayudan a determinar los requisitos del sistema [30].

- **Generador de código:** Permite ofrecer una herramienta que brinda una base de cómo realizar el código correspondiente al proyecto en curso. Puede permitir la generación del esqueleto del programa o del programa completo, además tener en cuenta la posibilidad de modificar manualmente el código y sincronizar esa información con la CASE.
- **Generador de documentación:** El módulo generador de la documentación ofrece a la herramienta CASE la construcción de documentos relacionados con las especificaciones que se tienen contenidas en el repositorio central y que son necesitadas en el transcurso del proyecto [30].

Clasificación de las CASE: No existe una única clasificación de las herramientas CASE y, en ocasiones, es difícil incluirlas en una clase en común. Pero algunos autores están de acuerdo en que se podrían clasificar de acuerdo a [25][30]:

- Las plataformas que soportan.
- La funcionalidad y las fases del ciclo de vida del desarrollo de sistemas que abarca (ver Figura 1. Las CASE y el proceso de desarrollo de software).
- La arquitectura de las aplicaciones que produce.

Por el tipo de herramienta que se modelará y desarrollará se utiliza una clasificación dependiendo de la funcionalidad, así:

- Herramientas integradas, I-CASE (Integrated CASE, CASE integrado): abarcan todas las fases del ciclo de vida del desarrollo de sistemas. Son llamadas también CASE workbench.
- Herramientas de alto nivel, U-CASE (Upper CASE - CASE superior), orientadas a la automatización y soporte de las actividades desarrolladas durante las primeras fases del desarrollo: análisis y diseño.
- Herramientas de bajo nivel, L-CASE (Lower CASE - CASE inferior), dirigidas a las últimas fases del desarrollo: construcción e implantación.
- Juegos de herramientas o Tools-Case, son el tipo más simple de Herramientas CASE. Automatizan una fase dentro del ciclo de vida. Dentro de este grupo se encuentran las herramientas de reingeniería y otras orientadas a la fase de mantenimiento.

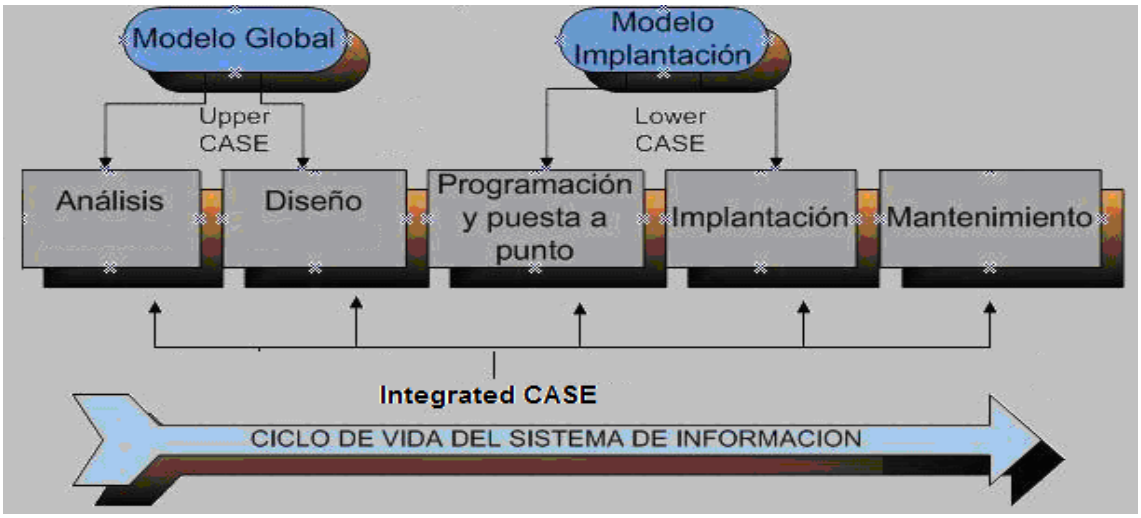

Figura 1. Las CASE y el proceso de desarrollo de software (Adaptado de [31]).

4 Descubrimiento del Conocimiento

El Descubrimiento de conocimiento en Bases de Datos (Knowledge Discovery in Database, KDD) es una amplia área que integra conceptos y métodos de diferentes disciplinas entre las que se puede incluir los campos de estadística, bases de datos, inteligencia artificial. Su objetivo es abstraer modelos de grandes bases de datos, y buscar nuevos, patrones no triviales y claves dentro del modelo abstraído, los patrones descubiertos deben ser validados con cierto grado de certeza, no deben ser obvios en el dominio del conocimiento y deben ser sustancialmente representativos para el usuario, para justificar el costo del proceso de KDD [32].

El término minería de datos (DM) se usa a menudo como un sinónimo para el proceso de KDD aunque estrictamente hablando es simplemente un paso dentro de KDD. La minería de datos usualmente usa modelos y algoritmos basados en búsqueda para encontrar patrones y modelos de interés. Las técnicas comúnmente usadas son árboles de decisión, programas genéticos, redes neuronales, programas de lógica inductiva, estadísticas bayesianas, optimización y otros semejantes. Los resultados de los algoritmos de minería se organizan y presentan al usuario de la forma más fácil posible para su mejor comprensión y uso [32].

Pasos de KDD: Se encuentran variaciones en los pasos del proceso de KDD. Algunas de estas variaciones no difieren su integridad, sino que algunas son más descriptivas que otras. Los nueve (9) pasos propuestos por Fayyad, Piatetsky-Shapiro y Smyth [34] son: entendimiento del dominio de la aplicación, creación de un dataset objetivo, limpieza y procesamiento de los datos, reducción y proyección de los datos, selección de las tareas de minería, selección de los algoritmos de minería, aplicación de la minería, interpretación de los modelos minados, y consolidación del conocimiento descubierto. Estos pasos se pueden resumir en seis (6) [32]: extracción de los datos, limpieza de los datos, ingeniería de los datos, ejecución de los algoritmos de minería, y

análisis de los resultados. En resumen los pasos se deben definir dependiendo del proyecto a desarrollar y el entorno (ver Figura 2. Proceso de KDD).

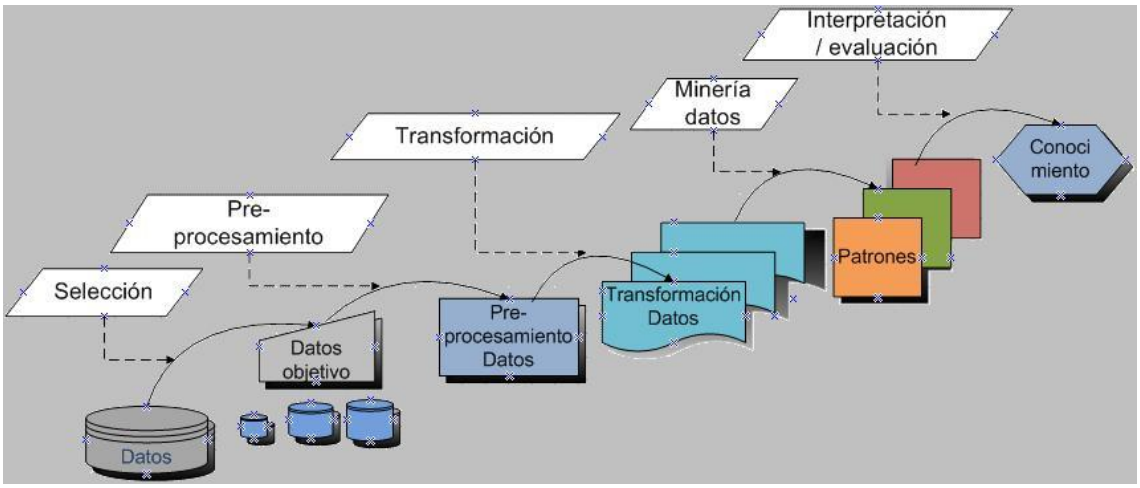

Figura 2. Proceso de KDD (Adaptado de [33]).

4.1 Metodologías y Procesos para el Desarrollo de Minería de Datos

Existen metodologías estándar para el proceso de minería de datos que pretenden facilitar la realización de nuevos proyectos con características similares, optimizar la planificación y dirección de los mismos, reducir su complejidad y permitir realizar un mejor seguimiento a estos [13], entre las que se destacan CRISP-DM (Cross – Industry Standard Process for Data Minning) [14] y SEMMA (Simple, Explore, Modify, Model, Assess) [15]. SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que CRISP-DM, mantiene como foco los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase donde SEMMA comienza realizando un muestreo de datos, mientras que CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista se puede considerar que CRISP-DM está más cercana al concepto real de un proyecto de minería de datos, pudiendo ser integrada con una Metodología de Gestión de proyectos específica que completaría las tareas administrativas y técnicas, además CRISP-DM es de libre distribución sin ningún costo a diferencia de SEMMA[13].

4.2 CRISP-DM: CRoss-Industry Standard Process for Data Mining

Es una iniciativa académica y empresarial, que define y valida un proceso estándar de minería de datos, de tal forma que se reduce el tiempo de ejecución y costo, y se mejora la administración de este tipo de proyectos [14].

La metodología CRISP-DM comprende una jerarquía de cuatro niveles. En el primer nivel se encuentran las fases que se componen de diferentes tareas genéricas las cuales son del segundo nivel, estas están diseñadas de tal forma que cubren todas las posibles situaciones que se presentan en el proceso de minería de datos. En el tercer

nivel están las tareas específicas que soportan los diferentes escenarios que puedan presentar las tareas genéricas, y en el cuarto nivel se encuentra la instancia del proceso, que describe las actividades específicas a realizar en un proyecto de minería de datos (ver Figura 3. Jerarquía de Fases y Tareas en CRISP-DM (Adaptado de [35])).

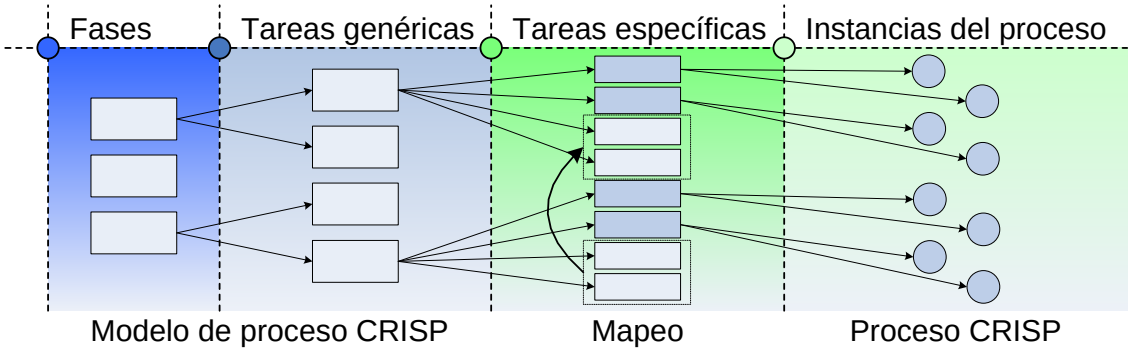

Figura 3. Jerarquía de Fases y Tareas en CRISP-DM (Adaptado de [35]).

El proceso CRISP-DM define una estructura para proyectos de minería de datos y suministra la orientación para su ejecución, consta de un modelo de referencia y una guía de usuario.

Modelo de referencia [35]: Este documento da una visión general del ciclo de vida de un proyecto de minería de datos, contiene las fases con sus objetivos, las tareas, las relaciones entre estas y las instrucciones pasó a paso que se deben llevar a cabo. Las fases definidas por el modelo de referencia son (ver Figura 4):

- **Business Understanding (Comprensión del negocio):** Busca comprender los objetivos del proyecto, los requerimientos del negocio, y luego convertir esta información en una definición de un problema. Incluye la realización de un plan preliminar.
- **Data Understanding (Comprensión de los datos):** Se hace una colección de datos, se familiariza con ellos, se identifican los problemas de calidad, se realizan las primeras percepciones o se detectan subconjuntos interesantes para formar las hipótesis.
- **Data Preparation (Preparación de los datos):** Cubre las actividades que construyen la estructura que alimentará la herramienta de modelado. Estas tareas son realizadas varias veces y en ningún orden preescrito. Se realiza la limpieza y transformación de los datos originales.
- **Modeling (Modelado):** Se seleccionan las técnicas a aplicar y se calibran sus parámetros a los valores óptimos. Algunas técnicas requieren que los datos posean formatos específicos, por esto, a menudo es necesario volver a la fase anterior (preparación de los datos).
- **Evaluation (Evaluación):** Se evalúan los modelos que parecen tener alta calidad, desde una perspectiva de análisis de datos. Se revisan los pasos realizados en la construcción del modelo, para cerciorarse que este cumple apropiadamente los objetivos del negocio y que no haya omitido algo importante. Para el fin de esta

fase, se debe tomar una decisión sobre el uso de los resultados de la minería de datos.

- **Deployment (Despliegue):** Se organizan y presentan los resultados de tal forma que el cliente pueda entender y acceder fácilmente a los mismos. Dependiendo de los requerimientos de esta fase puede ser simple, como generar un reporte, o compleja, como implementar un proceso de minería de datos repetible.

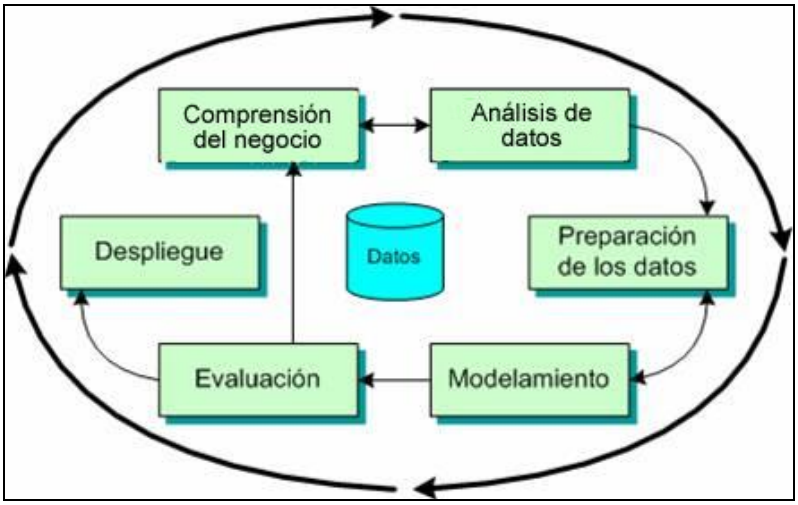

Figura 4. Fases del modelo de referencia CRISP-DM (Adaptado de [35]).

Guía de usuario [35]**:** La guía de usuario ofrece consejos más detallados, pistas por cada fase, y cada operación dentro de una fase, y ejemplifica cómo hacer un proyecto de minería de datos.

Esta guía de usuario es una excelente opción para desarrolladores de proyectos de minería de datos con poca experiencia y/o novatos.

Resultados de CRISP-DM [35]**:** CRISP-DM define resultados o reportes finales para cada Fase, involucrando en cada uno de estos Reportes las tareas y actividades más representativas de la Fase lo que permite resumir la información obtenida y documentar los resultados con el avance del Proyecto. Para dar un ejemplo en la Tabla 2 se presentan todos los resultados de CRISP-DM, para la Fase de Comprensión del Negocio se definen 12 resultados agrupados en las tareas Genéricas Determinación de los Objetivos del negocio, Evaluación de la Situación, Determinación de los objetivos de Minería de Datos y Construir el Plan del Proyecto. Pero el Reporte final de la Fase Comprensión del Negocio es resumido a seis (6) resultados[35]:

- Background: Provee una vista del contexto del Proyecto. Identifica en que área se ejecutara el Proyecto, el problema y el porque la minería de datos es una solución a este problema.
- Objetivos del Negocio y Criterios de Éxito: Describir los objetivos del Negocio y los objetivos del Proyecto en términos del Negocio. Para cada objetivo se definen los criterios de Éxito. Se debe explicar como se validará si el proyecto cumplió o no los objetivos.

- Inventario de Recursos: Permite identificar Personal, Fuentes de Datos, facilidades técnicas y otros recursos que se puedan necesitar en el Proyecto.
- Requerimientos, supuestos y restricciones: lista general de requerimientos del Proyecto que se espera obtener, se crean supuestos basado en la naturaleza del Problema y un listado de restricciones del Proyecto.
- Riesgos y contingencias: Identifica Problemas que se pueden presentar en el Proyecto, describe las consecuencias, el estado y la acción a tomar para reducir el impacto del riesgo.
- Terminología: permite familiarizarse con los problemas y unificar significados de conceptos.

Es preciso mencionar que los resultados (cursiva) son tareas específicas que contienen Actividades o instancias del Proceso.

4.3 Mapeo del Proceso CRISP-DM a Procesos Específicos

El proceso CRISP-DM esta diseñado de forma que cubra todas las posibles situaciones que se puedan presentar en un proyecto de minería de datos en general. Un valioso planteamiento de CRISP-DM es la creación de plantillas del proceso CRISP-DM a través del mapeo del Proceso Genérico a un Proceso Específico para un contexto de minería de datos.

Los contextos se definen en CRSP-DM por [35]:
- El **Dominio de aplicación**, es el área en específico en el cual se desarrolla el proyecto de minería de datos.
- El **tipo de Problema de minería de datos**. Existen Tipos de Problemas de Minería de datos (ver 4.4), el contexto es definido por uno de estos Problemas Generales.
- Las **Herramientas y técnicas** de minería de datos que son aplicadas en el proyecto.

Existen dos tipos de mapeos [35]:
- **Mapeando para el presente**: Se realiza para desarrollar solo un Proyecto, tomando las actividades que se crean necesarias para el desarrollo del proyecto. Este es un mapeo denominado sencillo para (probablemente) solo un uso.
- **Mapeando para el Futuro**: Se genera un modelo para un contexto predefinido teniendo como base experiencias pasadas o especializaciones en el desarrollo de proyectos en un contexto en específico. De tal manera que el modelo resultante sirva para orientar proyectos de contextos similares, es ahí donde se denomina un modelo del proceso especializado en términos o basado en CRISP-DM.

El tipo de mapeo adecuado para su propósito depende de su contexto específico de minería de datos y de las necesidades de la organización.

Comprensión del Negocio	Comprensión del los Datos	Preparación del los Datos	Modelado	Evaluación	Despliegue
Determinar los Objetivos del Negocio ➢ *Background* ➢ *Objetivos del Negocio* ➢ *Criterios de Éxito del Negocio* **Evaluar la situación** ➢ *Inventario de Recursos* ➢ *Requerimientos, supuestos y Restricciones* ➢ *Riesgos y contingencias* ➢ *Terminología* ➢ *Costo beneficio* **Determinar los Objetivos de Minería de Datos** ➢ *Objetivos de Minería de Datos* ➢ *Criterios de Éxito de la minería de Datos* **Construir Plan del Proyecto** ➢ *Plan del Proyecto* ➢ *Evaluación inicial de Herramientas y Técnicas.*	**Recolección Inicial de Datos** ➢ *Reporte de la recolección Inicial de datos.* **Descripción de Datos** ➢ *Reporte de Descripción de Datos.* **Exploración de Datos** ➢ *Reporte de Exploración de Datos.* **Verificar la calidad de los Datos** ➢ *Reporte de Calidad de los Datos.*	**Data Set** ➢ *Descripción del Data Set* **Selección de Datos** ➢ *Racionalmente por Inclusión/Exclusión* **Limpieza de Datos** ➢ *Reporte de Limpieza de Datos* **Construcción de Datos** ➢ *Atributos Derivados* ➢ *Generar Registros* **Integrar Datos** ➢ *Integración de Datos* **Formato de Datos** ➢ *Reformateo de Datos*	**Seleccionar la Técnica de Modelado** ➢ *Técnica de modelado* ➢ *Supuestos del Modelamiento* **Generar el Diseño de Test** ➢ *Diseño del Test* **Construir el Modelo** ➢ *Configuración de parámetros* ➢ *Modelo* ➢ *Descripción del Modelo.* **Evaluación del Modelo** ➢ *Evaluación del Modelo* ➢ *Revisión de la configuración de Parámetros*	**Evaluar los Resultados** ➢ *Evaluar los resultados de Minería de Datos Vs. Los Criterios de Éxito del Negocio.* ➢ *Aprobar modelos* **Revisar el Proceso** ➢ *Revisión del proceso* **Determinar el siguiente Paso** ➢ *Lista de Posibles Acciones* ➢ *Decisión*	**Plan de Despliegue** ➢ *Plan de Despliegue* **Plan de Monitoreo y Mantenimiento** ➢ *Plan de monitoreo y mantenimiento* **Producir los Reportes Finales** ➢ *Reportes Finales* ➢ *Presentación Final* **Revisión del Proyecto** ➢ *Documentación de Experiencias*

Tabla 2. Tareas Genéricas (Negrita) y resultados (*cursiva*) del modelo de Referencia CRISP-DM (Adaptada de [35]).

La estrategia básica para mapear un modelo de proceso genérico a un nivel especializado es el mismo para los dos tipos de mapeos, y es la siguiente [35]:

- Analizar el contexto específico.
- Eliminar los ítems no aplicables a su contexto.
- Adicionar cualquier detalle propio de su contexto específico.
- Especializar (**ejemplificar**) el contenido genérico de acuerdo a las características concretas de su contexto.
- Renombrar el contenido genérico para proveer significados más explícitos en su contexto por el bien de la calidad.

4.4 Tipos de Problemas de Minería de Datos

En los proyectos de minería de datos es común encontrar combinaciones de los tipos de problemas de minería de datos, los cuales deben ser definidos en las primeras etapas del proyecto. La Descripción y resumen de datos, la Segmentación, la Descripción de Conceptos, la Clasificación, la Predicción y el análisis de dependencia son tipos de Problemas de minería de datos. A continuación se presenta una corta descripción de cada uno de ellos.

Descripción y Resumen de Datos: La descripción y el resumen de datos apuntan a la descripción concisa de las características de los datos, típicamente en forma elemental y agregada. Esto da al usuario una descripción de la estructura de los datos. A veces la descripción y resumen de los datos puede ser el objetivo de un proyecto de minería de datos. Esta clase de problema estaría en lo mas bajo de la escala de problemas de minería de datos, en casi todos los proyectos de minería de datos es un objetivo subordinado en el proceso, que se realiza en las primeras etapas. Básicamente en la exploración inicial del análisis de datos puede ayudar a los usuarios a entender la naturaleza de los datos y formar hipótesis potenciales de la información oculta. La estadística descriptiva simple y las técnicas de visualización proporcionan las primeras ideas sobre los datos. Por ejemplo, la distribución de clientes por edad y regiones geográficas sugiere que partes de un grupo de clientes necesita para ser dirigida para futuras estrategias de comercialización (marketing). [35] [36]

Segmentación (Clustering): La segmentación es la separación de los datos en subgrupos o clases significativas e interesantes. Todos los miembros de un subgrupo comparten características comunes. Por ejemplo, en el análisis de cesta de compras, uno podría definir los segmentos de cestas según los artículos que ellos contienen. La segmentación puede ser realizada a mano o semi-automáticamente. El analista puede suponer ciertos subgrupos como relevantes para la pregunta de negocio, basada sobre un conocimiento previo o sobre el resultado de la descripción y el resumen de datos. En adición, hay también técnicas automáticas de segmentación/agrupamiento (clustering) que pueden descubrir las estructuras antes insospechadas y ocultas en datos que permite la segmentación. La segmentación a veces puede ser un objetivo de minería de datos. Entonces la detección de segmentos sería el objetivo principal de un proyecto de minería de datos. Por ejemplo, todas las direcciones en áreas de código postal con la edad mas alta que el promedio y un ingreso podrían ser seleccionadas para enviar publicidad para seguro de clínica de ancianos. En ocasiones la segmentación es un paso hacia la solución de otros tipos de problema, siendo así el objetivo de la segmentación mantener el tamaño de los datos manejables o encontrar los subconjuntos de datos homogéneos los cuales serán más fáciles para analizar. Generalmente grandes conjuntos de datos

y características variadas son un obstáculo para encontrar patrones interesantes. [35][36]

Descripción de Conceptos: La descripción de conceptos es una descripción comprensible de conceptos o grupos (clases). El objetivo no es el desarrollo de robustos modelos de predicción, sino para ganar ideas. Por ejemplo, una empresa puede estar interesada en el estudio sobre sus clientes más leales y desleales. Se debe encontrar una descripción de concepto de estos conceptos (clientes leales y desleales) la compañía infiere que podría estar hecho para encontrar clientes leales o transformar clientes desleales a clientes leales. [35][36]

Clasificación: La clasificación asume un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a diferentes clases, la etiqueta de clase es un valor (simbólico) discreto y es conocido para cada objeto. El objetivo de la clasificación es construir modelos de clasificación (a veces llamados clasificadores), que asignen la etiqueta de clase correcta a objetos antes no vistos y sin etiquetar. Los modelos de clasificación sobre todo son usados para el modelado predictivo. Las etiquetas de clase pueden ser otorgadas por la información del usuario o ser derivadas de la segmentación (cada segmento es una clase). La clasificación es uno de los tipos de problemas más importantes de minería de datos tiene una amplia gama de aplicaciones. Muchos problemas de minería de datos pueden ser transformados a problemas de clasificación. Por ejemplo en la asignación de créditos, para evaluar el riesgo de acreditar a un cliente nuevo. Esto puede ser transformado a un problema de clasificación para crear dos clases, clientes buenos y clientes malos. Un modelo de clasificación puede ser generado de los datos de cliente existentes de acuerdo a su comportamiento crediticio. Este modelo de clasificación puede entonces ser usado para asignar a clientes nuevos, una de las dos clases y aceptarlo o rechazarlo. [35][36]

Predicción: La predicción también tiene una gran gama de aplicaciones y es muy similar a la clasificación. La única diferencia es que en la predicción el atributo objetivo (la clase) no es un atributo cualitativo discreto, sino uno continuo. El objetivo de la predicción esta en encontrar el valor numérico del atributo objetivo para objetos no vistos. En la literatura, este tipo de problema es a veces llamado regresión. Si la predicción trata con datos de serie tiempo, entonces a menudo lo llaman pronosticación. [35][36]

Análisis de Dependencia (Reglas de Asociación): El análisis de dependencia consiste en encontrar un modelo que describe dependencias significativas (o asociaciones) entre registros de datos o acontecimientos. Las dependencias pueden ser usadas para predecir el valor de un registro con la información de otros registros. Aunque las dependencias pueden ser usadas para el modelado predictivo son más usadas para la comprensión. Las asociaciones son un caso especial de dependencias que describen las afinidades de registros de datos (esto es, registros de datos o los acontecimientos que con frecuencia ocurren juntos). Un típico escenario de aplicación para asociaciones es el análisis de cestas de compra. Allí, una regla como "en el 30 por ciento de todas las compras, la cerveza y cacahuetes han sido comprados juntos" es un ejemplo típico para una asociación. Los algoritmos para detectar asociaciones son muy rápidos y producen muchas asociaciones. Seleccionar el más interesante es un desafío. [35][36]

Estos tipos de Problemas son atacados por las técnicas de minería de Datos, sin importar el contexto. Es decir existen proyectos con el mismo tipo de problema pero

en un diferente contexto (Datos). En la Solución del Problema cambia es la técnica o técnicas de minería aplicadas y sus configuraciones.

4.5 Visualización del Flujo de Trabajo en Minería de Datos

La visualización ha sido un tema de gran Importancia en la adopción de tecnologías, por ejemplo los computadores solo fueron bien adoptados hasta que empezaron a trabajar con el sistema de ventanas (Windows) el cual reemplazo la interfaz de comandos, la gran acogida de la Internet se dio cuando se crearon los Navegadores gráficos (Browsers). De igual manera se han creado interfaces de usuario grafico para desarrollar el proceso de extracción de conocimiento con minería de datos y de esta forma lograr un mayor entendimiento del proceso y su adopción. La visualización del flujo de trabajo juega un gran rol en la minería de datos presentando los pasos de minería de datos en un formato grafico, lo que permite la racionalización y orientación de un individuo a través del proceso de minería de datos. El uso de un diagrama de flujo del proceso posibilita que una persona cubra todos los pasos críticos del proceso y facilita la gestión de actividades de minería de datos, teniendo en cuenta el contexto del modelo que realiza, comprendiéndolo mejor mediante la imagen completa del proceso realizado [37].

Existen iniciativas e implementaciones de flujos de trabajo gráficos de minería de datos basados en partes del proceso CRISP-DM, facilitando el desarrollo de actividades comprendidas en este proceso, específicamente en las fases de Comprensión de los Datos, Preparación de los Datos y especialmente en la fase de Modelado.[37]

El concepto de Flujo de trabajo visual (visual work flow) permite definir todos los pasos como iconos o nodos, obteniendo al final un grafico con el Proceso completo de Minería de datos [37]. Por ejemplo en CRISP-DM en la Fase de Conocimiento de los datos se necesita realizar la recolección inicial de datos lo cual en un WORK FLOW de minería de datos indicaría tener una fuente de datos (ver Figura 5), que necesitaría una configuración básica donde se le establezca por ejemplo de donde obtiene los datos y como.

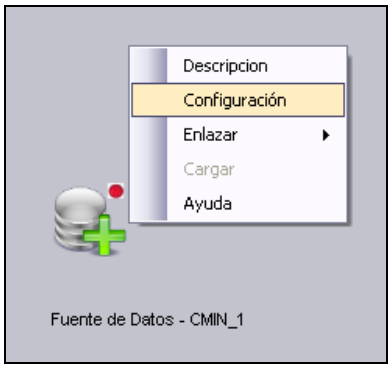

Figura 5. Work Flow de minería con Fuente de datos (Tomada de CMIN).

Entre las Tareas de la Fase de Procesamiento de los Datos se puede hacer reducción de atributos o cualesquier transformación de la Fuente de Datos, esto seria aplicar un filtro en el WORK FLOW a la Fuente de Datos (ver Figura 6. Fuente de Datos y Filtro de Columnas).

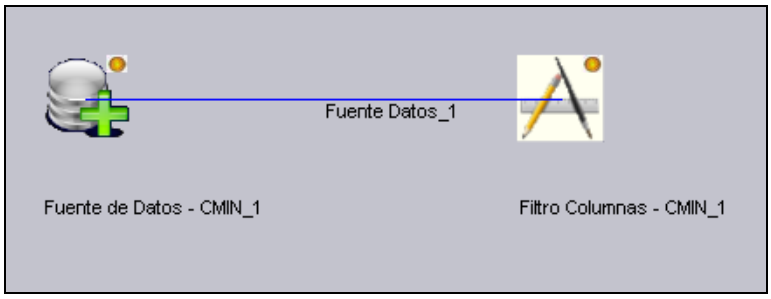

Figura 6. Fuente de Datos y Filtro de Columnas (Tomada de CMIN).

Terminando con el ejemplo, en la Fase de Modelado se selecciona una Técnica (por ejemplo un algoritmo de clasificación como K-nn) y se aplica a los datos para obtener los resultados (ver Figura 7. Proceso completo de minería de Datos.).

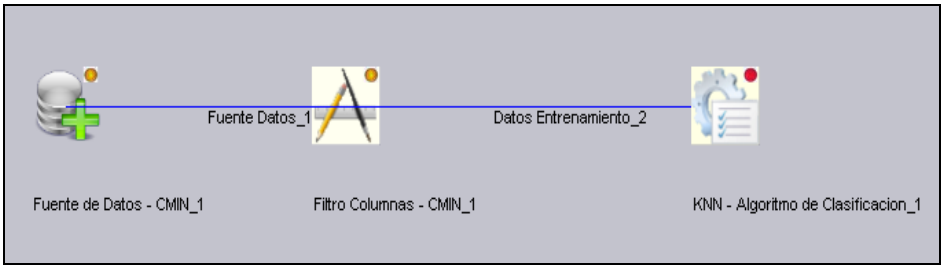

Figura 7. Proceso completo de minería de Datos. (Tomada de CMIN).

Al final se obtiene un modelo completo del Proceso de minería de datos en formato Grafico que especifica los pasos que se siguieron para obtener el resultado.

5 Aspectos Tecnológicos

Para el desarrollo (construcción) de la herramienta CASE, en especial la gestión dinámica y flexible de nuevos algoritmos de minería de datos que se incorporan en tiempo de ejecución dentro de la herramienta, se hizo necesario tener en cuenta tres conceptos fundamentales: las interfaces de software, el manejo de la Reflexión y los servicios Web XML. A continuación se revisan en forma general cada uno de ellos.

5.1 Interfaces de Software (Interface)

Las interfaces son la definición de un conjunto de miembros como propiedades, métodos y eventos que las clases pueden implementar. Las Interfaces solo definen las Firmas (declaraciones o contratos) de los miembros, pero no la funcionalidad interna de ellos. Esta funcionalidad es suplida por cada Clase o estructura que implemente la Interfaz [38].

Para citar un ejemplo de la utilidad de las Interfaces utilizaremos un contexto de minería de datos. Contemplando el Problema de Clasificación en minería de datos, se puede definir una interfaz (ver Tabla 3. Código C# de ejemplo de Definición de Interface) con los métodos que generalmente implementan las técnicas o algoritmos para solucionar este tipo de problema.

```
public interface Classification
{

 /// <summary>
 /// Metodo llamado Classify que Retorna un cadena
 /// y recive un entero n
 /// </summary>
 string Classify(int n);

 /// <summary>
 /// Metodo llamado Configuration que Retorna un entero
 /// y recive una cadena s
 /// </summary>
 int Configuration(string s);

}
```

Tabla 3. Código C# de ejemplo de Definición de Interface

Esta Interfaz podría ser una base o firma que los desarrolladores utilizan para crear algoritmos que solucionen el problema de clasificación, implementando los métodos de dicha interfaz (ver Tabla 4. Código C# de ejemplo de Implementación de la Interface).

```
Public class ALG_Clasification_1 : Classification
{
 /// <summary>
 /// Metodo que implementa el metodó Classify definido en Classification
 /// </summary>
 Public string Classify(int n)
 {
 ...
 Return "File" + n.toString();
 }
 /// <summary>
 ///Metodo que implementa el metodó Configuration definido en Classification
 /// </summary>
 Public int Configuration(string s)
 {
 ...
 Return intNIteraciones;
 }
}
```

Tabla 4. Código C# de ejemplo de Implementación de la Interface

5.2 Reflexión (System.Reflection)

En .NET los ensamblados están compuestos por módulos y estos a su vez por TYPES (Clases, Delegados o Enumeraciones), los TYPES contienen propiedades, eventos, métodos, campos, etc. En la Tabla 5 se presenta como cargar un Assembly desde la ruta de un archivo con extensión ".dll" y luego consultar las Clases que contiene. La tecnología reflexión permite obtener información del ensamblado y sus TYPES, con esta información se puede verificar que métodos contiene un TYPE, el nombre de cada método, cuantos parámetros recibe y de que tipo, que retorna y si el método es estático o no [39].

```
using System.Reflection;

 // Se carga el Assembly
 oNewAssemblie = Assembly.LoadFrom(strPathFileTo_dll);
 // Si tien Types
 if (oNewAssemblie.GetTypes().Length > 0)
 {
 int i = 0;
 // Declaramos un listado de strings
 strNamesPaces = new string[oNewAssemblie.GetTypes().Length];
 // recorremos los Types del Assembly
 foreach (Type oNamespace in oNewAssemblie.GetTypes())
 {
 // llenamos la lista
 strNamesPaces[i] = oNamespace.FullName;
 i++;
 }
 }
```

Tabla 5. Código C# para cargar Assembly y obtener lista de Types y NamesPaces (Tomado del fuente de CMIN).

Continuando con el Ejemplo citado en Interfaces de Software (Interface), con Reflexión se puede verificar si un ensamblado contiene una clase que implemente o cumpla con la interfaz software definida en la Tabla 3 [38]. Si se genera una dll de la interfaz Classification y otra dll del algoritmo que implementa esta interfaz ALG_Clasification_1, se puede cargar los Assemblies y los Types (clases, Delegados y Estructuras) de estas dll, como se mostró en la Tabla 5, para comparar la información de los Assemblies mediante Reflexión (ver Tabla 6).

```
public static bool ValidarTypesDeAssemblies(Type TypeClasification, Type TypeALG_1)
{ // Obtenemos los Metodos de los Types
 MethodInfo[] MetodosTypeClasification = TypeClasification.GetMethods();
 MethodInfo[] MetodosTypeALG_1 = TypeALG_1.GetMethods();
 bool blnIquals = false;
 // recorremos los Metodos de la Interface Classification
 foreach (MethodInfo mTypeClasification in MetodosTypeClasification)
 {
 blnIquals = false;
 // Bucamos cada metodo definido en la Interface en los métodos del ALG_1
 foreach (MethodInfo mTypeALG_1 in MetodosTypeALG_1)
 { // Compara los método
 if (ValidarMetodos(mTypeClasification, mTypeALG_1))
 { // si son iguales
 blnIquals = true;
 }
 }
 if (!blnIquals)
 { // Si no encuentra alguno de los métodos de la Interface no Cumple con ella
 return false;
 }
 }// Todos los métodos de la Interface se encontraron en ALG1 y son iguales en su
 //definición
 return true;
}

private static bool ValidarMetodos(MethodInfo m1, MethodInfo m2)
{ //Compara los nombres
 if (m1.Name == m2.Name)
 { //Compara el tipo de Retorno
 if (m1.ReturnType == m2.ReturnType)
 { //Carga los Parametros de los Métodos
 ParameterInfo[] ParraInfom1 = m1.GetParameters();
 ParameterInfo[] ParraInfom2 = m2.GetParameters();
 bool IsEquals = false;
 // Compara Si tienen el mismo numero de Parametros
 if (ParraInfom1.Length == ParraInfom2.Length)
```

```
 {
 for (int i = 0; i < ParraInfom1.Length; i++)
 { // Compara si los parametros son del Mismo Tipo
 if (ParraInfom1[i].ParameterType == ParraInfom2[i].ParameterType)
 {
 IsEquals = true;
 }else{
 return false;
 }
 }
 if (ParraInfom1.Length == 0)
 IsEquals = true;
 return IsEquals;
 }
 }
 }
 // Los métodos son iguales
 return false;
}
```

Tabla 6. Código C# Comparación de Assemblies (Tomado del fuente de CMIN).

Dado que el ALG_Clasification_1 cumple con la Interfaz Classification con reflexión se puede en tiempo de ejecución cargar el Assembly de ALG_Clasification_1 y crear instancias de los TYPES (en este caso la Clase ALG_Clasification_1 definida en Tabla 4) e invocar los métodos de la instancia, soportado con la información que Provee las definiciones de la Interfaz Classification (ver Tabla 7).

```
public object EjecutarMetodo()
{ // Se carga el Type del Assembly
 Type oTypeALG = CargarAssemblyType("C:/ALG_Classification_1.dll", "ALG_Clasification_1");
 Object instance = null;
 // Se carga la Informacion del Metodo Configuration
 MethodInfo metodo = oTypeALG.GetMethod("Configuration");
 // Si no es un método estatico
 if (!metodo.IsStatic)
 { // Se crea una instancia de la Clase ALG_Clasification_1
 instance = Activator.CreateInstance(oTypeALG);
 }
 // Creamos los argumentos para invocar el Método Configuration
 // Por la Interfaz sabemos que es un Parametro y es tipo string
 object[] args = new object[1];
 args[0] = "X=1; Y=2; Z=3";
 // Invocamos el Método de la dll cargada de C:/ALG_Classification_1.dll
 return metodo.Invoke(instance, args);
}


public Type CargarAssemblyType(string Path, string NamesPace)
{ // Carga el Assembly
 Assembly oAssembly = Assembly.LoadFrom(Path);
 // Carga el Type del Assembly
 return oAssembly.GetType(NamesPace, true);
}
```

Tabla 7. Código C# Invocación de métodos de un Assembly en ejecución. (Tomado del Fuente de CMIN).

5.3 Servicios Web XML

Un servicio Web XML es un componente de software que se comunica con otras aplicaciones codificando los mensaje en XML y enviando estos mensaje a través de protocolos estándares de Internet tales como el Hypertext Transfer Protocol (HTTP). Intuitivamente un Servicio Web es similar a un sitio Web que no cuenta con un

interfaz de usuario y que da servicio a las aplicaciones en vez de a las personas. Un Servicio Web XML, en lugar de obtener solicitudes desde un navegador y retornar páginas Web como respuesta, lo que hace es recibir solicitudes a través de un mensaje formateado en XML desde una aplicación, realiza una tarea y devuelve un mensaje de respuesta también formateado en XML [40]. Mediante los servicios Web las aplicaciones pueden comunicarse e intercambiar funcionalidad e Información entre si.

Para este Proyecto fue esencial tener en cuenta esta tecnología en el Proceso de Actualización de procesos definidos en CMIN (CRISP-DM Versión 1, 2, etc.) y de esta manera permitir que los usuarios de CMIN puedan compartir nuevos procesos y además algoritmos entre ellos. De esta manera se puede contar con un servidor central, que expone los Servicios Web de **Actualización de Procesos** que permiten al usuario cliente actualizar los Procesos de su aplicativo CMIN local y **Actualización de Algoritmos** mediante el cual descargara algoritmos del servidor, que han sido Probados y validados por otros usuarios de la comunidad CMIN.

Es de Notar que las aplicaciones de los Usuarios llamados Cliente cuentan por si solas con la capacidad de extender tanto los Procesos como los Algoritmos, lo descrito aquí hace referencia a un trabajo futuro que es posible, gracias al actual diseño de la aplicación y que posibilita la creación de una comunidad de usuarios alrededor de CMIN.

PARTE 3: MODELO DE CMIN Y MODULOS

6 METODOLOGÍA

El enfoque metodológico del proyecto está guiado por una instanciación del Proceso Unificado (Unified Process, UP), estructurado en tres (3) ciclos de desarrollo iterativos e incrementales, teniendo en cuenta las siguientes fases [41]: Planeación y Elaboración, Construcción, y Transición.

6.1 PLANEACIÓN Y ELABORACIÓN

En esta fase se obtuvo una comprensión del contexto en el que se iba a desarrollar CMIN teniendo en cuenta que se realizaron actividades de conocimiento y entendimiento del tema de minería de datos y de la metodología CRISP-DM. Además se obtuvo una lista inicial de los requerimientos del sistema desde el punto de vista funcional, como los relacionados con la distribución de los datos. Los artefactos necesarios para completar esta fase fueron: casos de uso de alto nivel, el modelo conceptual preliminar, el diagrama de secuencia preliminar, el glosario (preliminar), un modelo de distribución, y una arquitectura preliminar del sistema.

6.2 CONSTRUCCIÓN

A partir de las actividades y productos obtenidos en la planeación y elaboración se inició esta fase, cuyo propósito fue obtener la versión operativa inicial del sistema, con la calidad y los requisitos necesarios para su aplicación. Esta etapa se dividió en tres (3) ciclos de desarrollo, donde cada uno contuvo las siguientes etapas:

- **Análisis:** Esta fase se basó en los requerimientos obtenidos en la fase de planeación y elaboración, y partiendo de los requerimientos específicos del sistema, se obtuvo una concepción clara de los requisitos a desarrollar en cada ciclo, se desarrolló la descripción de los casos de uso de alto nivel y el modelo conceptual específico para cada fase.

- **Diseño:** Teniendo en cuenta el análisis del sistema se procedió a generar una solución lógica del prototipo software que finalmente fue implementada, para lo cual se diseñaron los casos de uso reales, los diagramas de clases de diseño, el diagrama de persistencia de la base de datos, los diagramas de secuencia, de paquetes y de despliegue.

- **Implementación:** Se implementaron los componentes lógicos obtenidos en la etapa de diseño, se documentó el código, y se realizó un conjunto de pruebas preliminares de caja blanca y negra.

- **Pruebas:** Se intercambiaron los productos desarrollados por los autores y se realizó una revisión y prueba detallada. Además se definieron un conjunto de valores de prueba y se realizaron formalmente las pruebas de caja blanca y caja negra con el fin de verificar el resultado de la implementación generada en esta fase.

6.3 TRANSICIÓN

Luego de culminar la fase de construcción del sistema, se realizaron las respectivas pruebas de caja negra, de integración y del sistema en su totalidad, con el fin de realizar su adecuación final e implantación. Se realizo la evaluación de la aceptación del sistema por parte de usuarios reales, en este caso, estudiantes de ingeniería de

sistemas, que estén desarrollando trabajos de grado en minería de datos y/o ingenieros interesados en la temática. Se evaluó principalmente cuatro ítems, la funcionalidad con respecto a lo planteado por CRISP-DM, la usabilidad, el apoyo al proceso de aprendizaje que genera CMIN al usuario y finalmente, la utilización de los algoritmos de minería disponibles y la posibilidad de ser ampliados.

- **Diseño de la Prueba:** Con el fin de cubrir todos estos ítems se diseño una prueba en la que los usuarios realizan un test inicial de conocimiento, seguido de la utilización de CMIN realizando un taller de minería de datos, donde posteriormente se les muestre la posibilidad de adicionar nuevos algoritmos, explicando el proceso para ello. Luego teniendo en cuenta que para cada actividad de un proyecto se presenta información detallada, se aplica al grupo el mismo test inicial para verificar si se apoyo el crecimiento de aprendizaje sobre la metodología CRISP-DM, finalmente se aplica un test de usabilidad para la herramienta.

6.4 DOCUMENTACIÓN Y DIVULGACIÓN DE RESULTADOS

Para la fase de divulgación se creo un Artículo que resume el contenido de la presente monografía que será presentado en el IEEE International Conference on Data Mining series (ICDM 2009) a realizarse en Miami, Florida del 6-9 de diciembre de 2009 y para el cual se debe entregar la versión definitiva en ingles antes del 26 de Junio de los corrientes. Finalmente esta monografía y la sustentación de la tesis ante los jurados definidos por la FIET hacen parte de esta fase.

7 MODELO CMIN 1.0

Para comprender mejor el funcionamiento de CMIN 1.0, a continuación se presenta el modelo conceptual en el cual se representan los conceptos que hacen parte del sistema, así como las relaciones existentes entre estos, ver en la Figura 8. Modelo Conceptual de CMIN 1.0

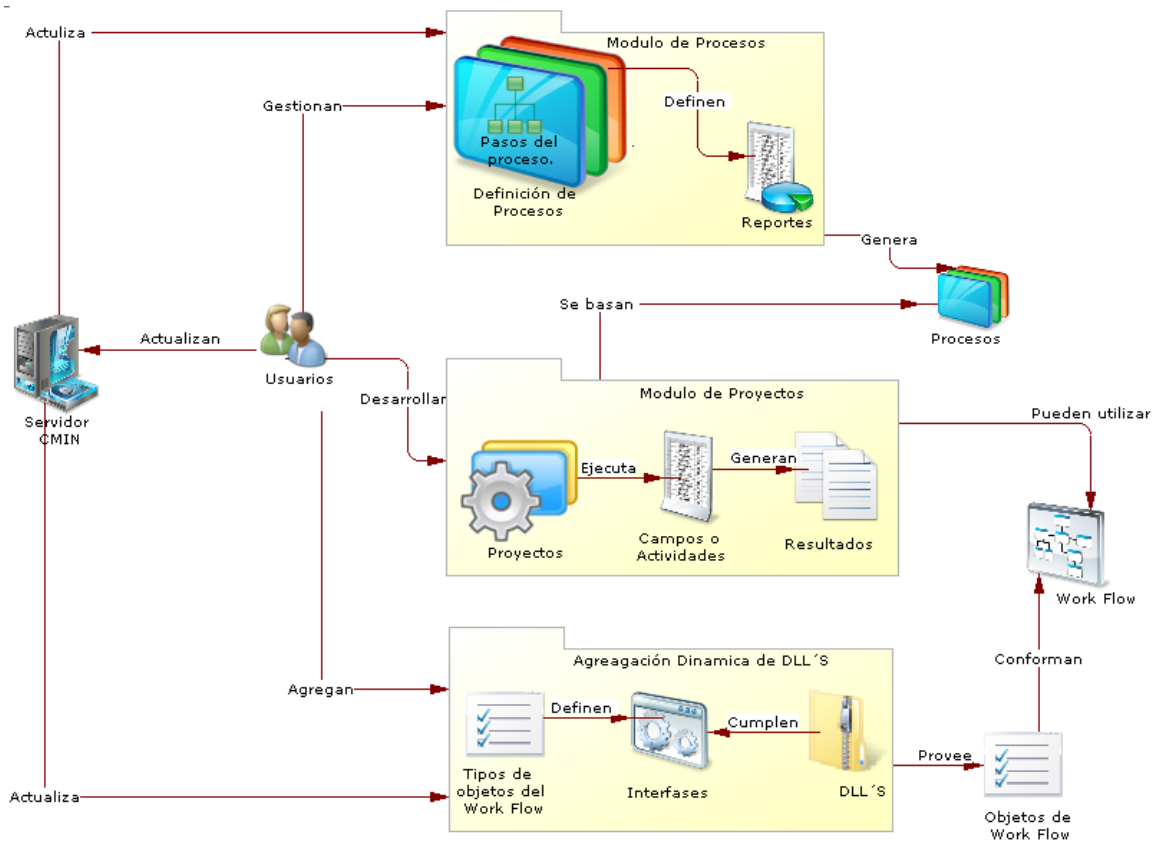

Figura 8. Modelo Conceptual de CMIN 1.0

A continuación se explica en forma breve aquellos conceptos y relaciones del modelo de CMIN 1.0 que permiten percibir, identificar y describir su funcionamiento.

- **Usuarios**: Representa a las personas que pueden utilizar el sistema, los cuales pueden ser principiantes o expertos en minería de datos lo que significa que la única diferencia entre los usuarios es el nivel de conocimiento en el tema, pero podrían llegar a realizar las mismas actividades en CMIN 1.0.

- **Modulo de Procesos**: Representa el modulo que permite la agregación de procesos o metodologías, definiendo la estructura, pasos y entregables propuestos por cada una. Este modulo contiene los siguientes conceptos para su funcionamiento:

  - **Definición de procesos**: Representa la acción de construir un proceso o metodología mediante la agregación y definición de sus pasos, campos o actividades que propone para el desarrollo de un proyecto de minería de datos.
  - **Reportes**: Representa los documentos o entregables que se deben arrojar durante un proyecto, los cuales servirán como soporte durante la ejecución del mismo.

- **Procesos:** Representa los procesos o metodologías que se han agregado a CMIN 1.0 y que servirán como base para los proyectos de minería a desarrollar.

- **Modulo de Proyectos**: Representa el modulo que permite la ejecución de un proyecto de minería de datos basado en uno de los procesos o metodologías agregado en el modulo de procesos. Este modulo contiene los siguientes conceptos para su funcionamiento:

  - **Proyectos**: Representa el conjunto de proyectos que se han creado en CMIN 1.0 y están en curso o terminados.
  - **Campos o Actividades**: Representan los campos que pertenecen a un paso, los cuales especifican las actividades que se deben realizar para cumplir con el fin del paso al que pertenecen.
  - **Resultados**: Representan el resultado de la realización de una actividad, el cual puede ser: una sugerencia, un texto explicativo o una plantilla que se debe diligenciar.

- **Agregación dinámica de Librerías de Enlace Dinámico (DLL's)**: Representa el modulo que permite la agregación de objetos que servirán para la ejecución del Work Flow, por medio de DLL's con su funcionalidad. Este modulo contiene los siguientes conceptos para su funcionamiento:

  - **Tipos de objetos del Flujo de Trabajo (Work Flow, WF)**: Representa el conjunto de tipos de objetos reconocidos por CMIN 1.0 para ser agregados y posteriormente ser utilizados por el WF ofrecido por CMIN 1.0.
  - **Interfases**: Representa el conjunto de interfases que deben cumplir las DLL's para poder ser agregadas al conjunto de objetos que serán utilizados por el WF.
  - **DLL's**: Representan el conjunto de DLL's que poseen la implementación de los objetos del WF.

- **Objetos de WF**: Representa el conjunto de objetos que se han agregado a CMIN 1.0 y podrán ser utilizados por el WF, el cual puede crecer a medida que los usuarios realicen implementaciones nuevas de cualquiera de los tipos de objetos del WF especificados en CMIN 1.0.

- **Work Flow**: Representa la herramienta que ofrece CMIN 1.0 para que los usuarios realicen actividades propias de minería, y puedan crear modelos utilizando la tareas de minería propuestas por CMIN 1.0.

- **Servidor CMIN**: Representa el servidor que aloja nuevas definiciones de metodologías, así como nuevas implementaciones de objetos del WF por medio de DLL's, para que los usuarios actualicen CMIN 1.0 si así lo requieren ya que CMIN 1.0 se ejecuta independientemente de este servidor.

Con la combinación de cada uno de estos conceptos y sus relaciones, el presente modelo conceptual formula como es el flujo normal del funcionamiento de la herramienta CMIN 1.0 para el soporte en el desarrollo de proyectos de minería de datos.

8 ARQUITECTURA DEL SISTEMA

El modelo general de la aplicación se realiza a través de una arquitectura de tres capas (ver Figura 9. Diagrama de la arquitectura del ambiente), la cual es muy útil en la construcción de la aplicación, debido a que permite dividir responsabilidades y además permite diminuir complejidad. Las tres capas de la arquitectura son:

- **Capa de Interfaz**: componentes de interfaz de usuario.
- **Capa de Negocio**: entidades del negocio las cuales componen los objetos del negocio que representan conceptos del dominio del problema, cumpliendo con los requisitos de la aplicación.
- **Capa de Servicios**: objetos que ofrecen servicios para soporte, como el acceso a datos.

En cada capa se definen componentes necesarios para un óptimo funcionamiento. Es de notar que el funcionamiento de CMIN 1.0 normalmente cumple con la arquitectura del lado del cliente, además existe un componente de servicios web que permite el comportamiento como cliente servidor en las ocasiones que el usuario quiera actualizar la herramienta por medio del servidor. El cual solo debe ser manejado por un usuario experto quien hace la definición de nuevos procesos estándar y se encarga de validar si los resultados de los nuevos algoritmos son correctos, para su difusión.

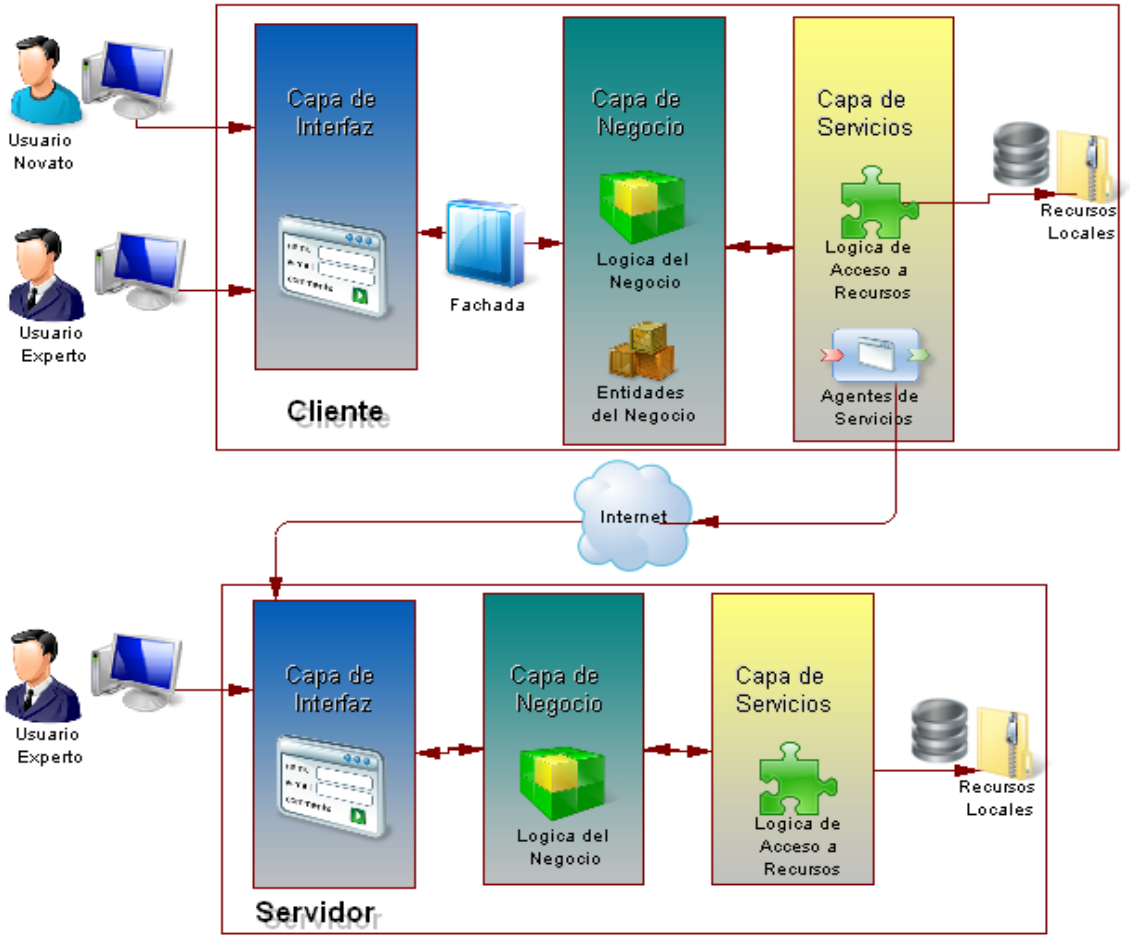

Figura 9. Diagrama de la arquitectura del ambiente.

8.1 Capa de Interfaz.

La capa de interfaz expone los servicios de la capa de lógica de negocios a los usuarios. Esta interpreta y procesa las peticiones del cliente, además puede interactuar con la capa de lógica de negocios.

La capa de interfaz de la aplicación está organizada lógicamente en paquetes, conteniendo los elementos mencionados. (Ver Figura 10. Capa de Interfaz).

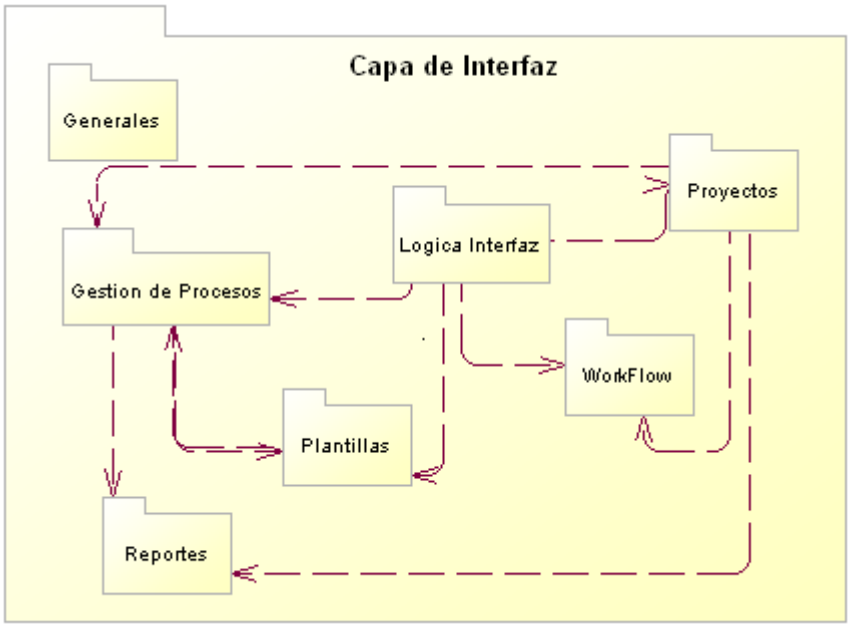

Figura 10. Capa de Interfaz.

El paquete **Generales**, son formas Windows que ejecutan las opciones iniciales del sistema como: ingreso al sistema, configuración del servidor de bases de datos y el formulario principal que actúa como contenedor de interfaz.

El paquete **Lógica Interfaz** contiene los objetos propios para realizar el manejo de la capa de interfaz.

El paquete **Gestión de Procesos** contiene las interfaces que permiten la gestión de procesos, así como de sus pasos y la definición de sus reportes.

El paquete **Proyectos** contiene las interfaces que permiten el desarrollo de un proyecto.

El paquete **WorkFlow** contiene las interfaces que permiten la utilización del work flow que permite realizar actividades propias de minería de datos.

El paquete **Plantillas** contiene las interfaces que permiten la gestión de plantillas, definir los pasos que se deben utilizar y los reportes a utilizar.

El paquete **Reportes** contiene las interfaces que permiten generar y visualizar los reportes del proyecto y los reportes de Gantt para el seguimiento de un proyecto.

8.2 Capa de Lógica de Negocio

La capa de lógica de negocio contiene los objetos y servicios de negocio de la aplicación. Recibe peticiones de la capa de interfaz, procesa la lógica de negocio basada en las peticiones, y media en los accesos a los recursos de la capa de lógica de servicios.

Esta capa contiene la clase Fachada por medio de la cual se implementa el patrón fachada para controlar la interacción entre la capa de interfaz y la capa de negocio. Además del patrón fachada se utilizaron los siguientes patrones: el patrón Singleton, por medio del cual permite la instancia de un objeto accesible globalmente, y que es único para realizar el manejo de la clase fachada y solo se tuviera una sola instancia evitando saturar la memoria con varias instancias de esta; internamente esta implementado el patrón Modelo Vista Controlador ya que los objetos del negocio (modelo) están separados de las interfaces (vista) por medio de las capas de negocio y de interfaz, para ejecutar las respuestas de las peticiones de los usuarios se realiza el manejo de los eventos (controlador) para ejecutar los procesos necesarios.

La funcionalidad de está capa se basa en entidades del negocio, las cuales son objetos que representan las tablas existentes en la base de datos facilitando el desarrollo de esta capa, para crear las entidades se utiliza el componente **Subsonic** el cual es una herramienta que se integra con Visual Studio 2005 por medio de la cual se pueden crear las entidades del negocio y además facilita el acceso a los datos. Los objetos del negocio que implementaran las reglas del negocio, se crean basándose en las entidades del negocio. (Ver Figura 11. Capa de Lógica de Negocio).

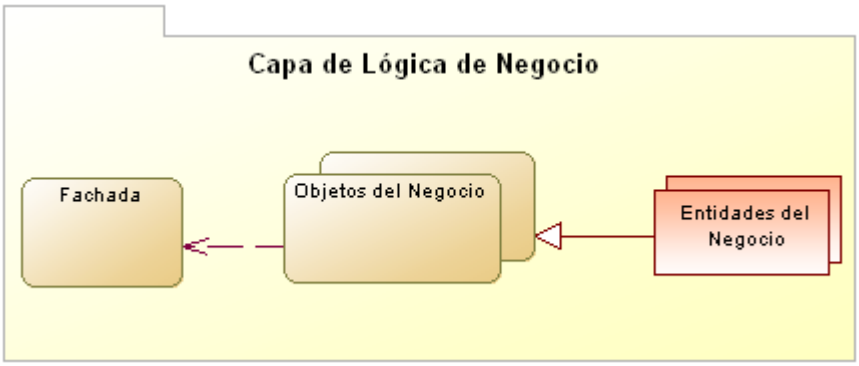

Figura 11. Capa de Lógica de Negocio

8.3 Capa de Lógica de Servicios

La capa de lógica de servicios representa la forma en la cual se manipula y persiste la información, es decir, esta capa se encarga de recibir solicitudes de almacenamiento o recuperación de información desde la capa lógica del negocio a través de diferentes métodos, el manejo de los datos se realiza por medio del componente **Subsonic** el cual facilita el manejo del acceso a los datos. (Ver Figura 12. Capa de Lógica de Servicios).

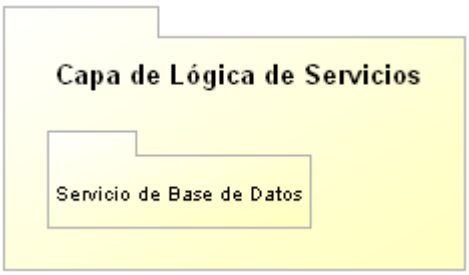

Figura 12. Capa de Lógica de Servicios.

9 CMIN: Sistematización de CRISP-DM

El principal objetivo de este trabajo fue Desarrollar una herramienta CASE integrada para soportar y orientar el desarrollo de proyectos de minería de datos, Basada en una Metodología estándar como lo es CRISP-DM. Este objetivo origino la Creación del Modulo de Gestión de Procesos, que permite Editar los Procesos del sistema, y de cada Proceso sus Pasos y de cada Paso sus Actividades o Campos. Este Modulo resuelve el Problema de Actualización de futuras versiones de CRISP-DM. En la actualidad CMIN tiene creado el Proceso CRISP-DM V1.0 el cual contiene todos los pasos y actividades planteadas en la metodología CRISP-DM versión 1.0. Este modulo posibilita la adición de nuevos procesos para cada versión de CRISP-DM.

9.1 Modelo para el Modulo de Gestión de Procesos

Para que CMIN soporte varios Procesos, se diseño un modelo de base de Datos que contempla el almacenamiento del Proceso, su jerarquía de pasos y las actividades o campos de los Pasos. Para explicar el modelo de base de datos que soporta la Gestión de Procesos se utilizaran las etiquetas de las relaciones ("**Relación (1…7)**") que se presentan en la Figura 13. Modelo de Base de Datos de Gestión de Procesos.

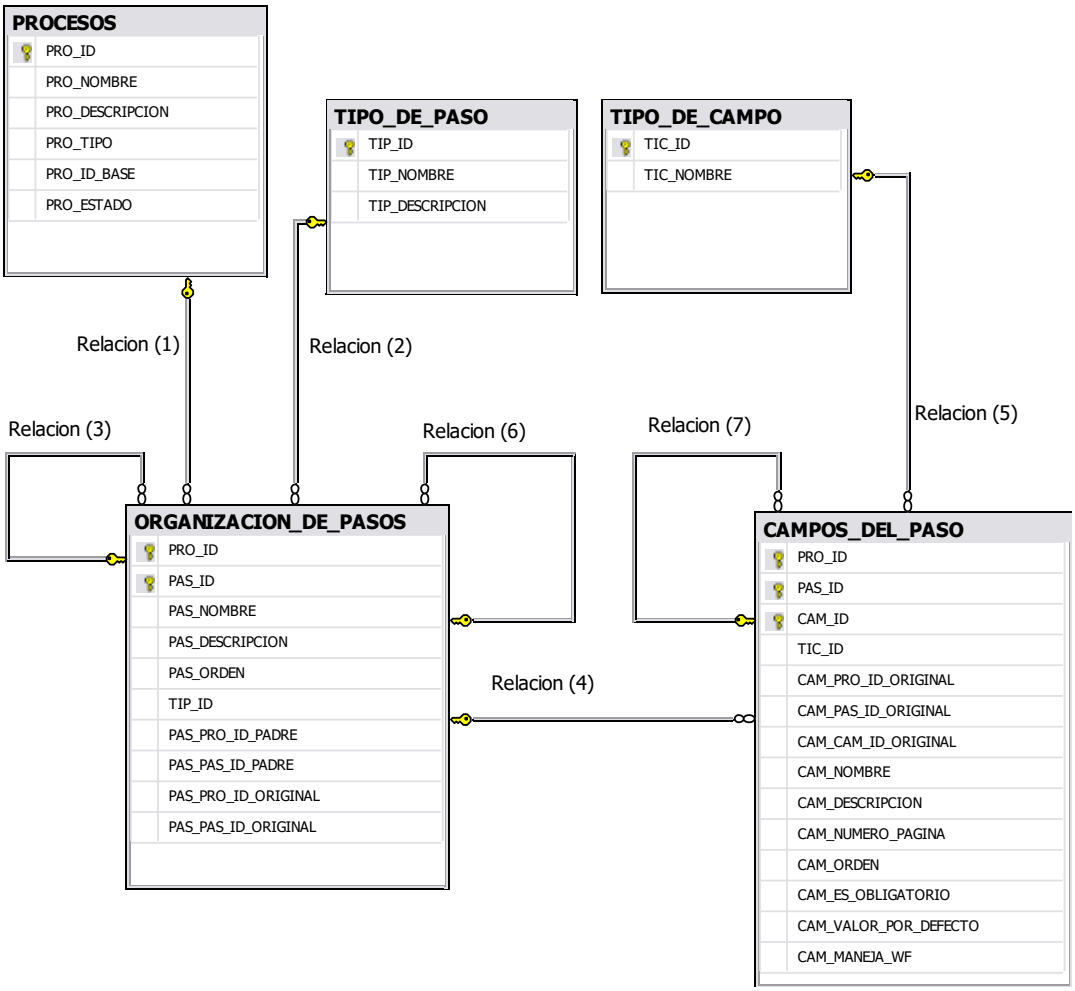

Figura 13. Modelo de Base de Datos de Gestión de Procesos.

La Información básica del Proceso se registra en la Tabla **PROCESOS**, los pasos del Proceso en la tabla **ORGANIZACION_DE_PASOS**, la **Relación (1)** establece a que Proceso pertenece cada Paso. Para cumplir con la jerarquía Planteada en CRISP-DM (ver Figura 3), definida por la Metodología en el Nivel 1, las Fases en el Nivel 2, las Tareas Genéricas en el Nivel 3 y las Tareas Especificas en el Nivel 4, se creo la tabla **TIPO_DE_PASO** en donde se registran el nombre y la definición de cada nivel de la jerarquía, por tanto la **Relación (2)** Indica de que tipo es el Paso, si es una FASE, Tarea Genérica ó Tarea Especifica, etc.

La relación recursiva **Relación (3)** permite establecer Pasos hijos y es la referencia a el Paso Padre del Paso. Cuando se crea un Proceso en CMIN el sistema automáticamente crea un Paso de tipo Metodología el cual no tiene Padre por tanto su **Relación (3)** referencia a vacío (Nulo). El Modulo de Gestión de Procesos permite Adicionar a el paso de Tipo metodología los Pasos del Siguiente Nivel y realizar lo mismo para cada uno de sus pasos hijos (ver Figura 16, Figura 17 y Figura 18).

El último Nivel de la jerarquía CRISP-DM denominado "Instancias del Proceso" son las Actividades propuestas a realizar en un Proyecto de Minería de Datos y este Nivel **NO** se registra en la Tabla **TIPO_DE_PASO** por tanto "Instancias del Proceso" **NO** es un "tipo de Paso" en CMIN. Teniendo en cuenta esto podemos afirmar que los Pasos son de un Nivel superior al Nivel "Instancias del Proceso" en consecuente los

Pasos contienen actividades o campos definidos en este Nivel. Las "Instancias del Proceso" son almacenadas en la tabla **CAMPOS_DEL_PASO** y la **Relación (4)** Identifica a que paso pertenecen los Campos o Actividades. En esta Tabla además de las actividades también se almacenan las **sugerencias** que presenta la metodología CRISP-DM en la guía de usuario (ver CRISP-DM: CRoss-Industry Standard Process for Data Mining), teniendo así Campos del paso de tipo sugerencia y Actividad. Estos tipos de Campos son almacenados en la Tabla **TIPO_DE_CAMPO** y la **Relación (5)** establece el tipo del Campo o Actividad. Este Modulo permite Editar los Campos o Actividades de un Paso (Ver Figura 19).

CRISP-DM plantea mapear el Proceso Genérico a un Proceso específico para un contexto de minería de datos (ver Mapeo del Proceso CRISP-DM a Procesos Específicos), este modulo de Gestión de Procesos permite la creación de **Plantillas del Proceso** (Procesos Específicos) para ser utilizadas al igual que los procesos como base para el desarrollo de Proyectos.

Para la creación de una Plantilla de Proceso en CMIN (ver Figura 20) se debe seleccionar el Proceso Base con el que vamos a realizar la labor de mapeo. Las Plantillas de Procesos son almacenadas en la Tabla **PROCESOS** diferenciándose por el valor de la columna **PROCESOS.PRO_TIPO** y que el valor de la Columna **PROCESOS.PRO_ID_BASE** para las Plantillas es el Identificador (Guid) del Proceso BASE y para los Procesos es Vacío o Nulo. Cuando se crea la Plantilla la Información base se registra en la tabla **PROCESOS** creando un nuevo Identificador (Guid) para la Plantilla, una Rutina de este Modulo carga los pasos del Proceso Base en conjunto con sus Campos en memoria y cambia valores de los Pasos y los Campos en las siguientes Columnas:

En los Pasos.
- **ORGANIZACION_DE_PASOS.PRO_ID** por el valor del Identificador (Guid) de la Nueva Plantilla del Proceso.
- **ORGANIZACION_DE_PASOS.PAS_PRO_ID_ORIGINAL** y **ORGANIZACION_DE_PASOS.PAS_PAS_ID_ORIGINAL** (los cuales son vacíos o Nulos para los Pasos de PROCESOS) por el Identificador (Guid) del Proceso BASE y el Identificador del paso.

En los Campos:
- **CAMPOS_DEL_PASO.PRO_ID** por el valor del Identificador (Guid) de la Nueva Plantilla del Proceso.
- **CAMPOS_DEL_PASO.CAM_PRO_ID_ORIGINAL**, **CAMPOS_DEL_PASO.CAM_PAS_ID_ORIGINAL** y **CAMPOS_DEL_PASO. CAM_CAM_ID_ORIGINAL** (los cuales son vacíos o Nulos para los campos de PROCESOS) por el Identificador (Guid) del Proceso BASE, el Identificador del paso al que pertenece y el Identificador del campo.

Después de cambiar estos valores en memoria la rutina registra en la base de datos esta información generando una **copia** completa del Proceso BASE con el objetivo de que la labor de Mapeo en CMIN (ver Figura 21. Mapeo del Proceso BASE a un Proceso Especifico.) se realice sobre los pasos de la nueva Plantilla y de esta manera no afecte el Proceso BASE. La copia de los Pasos y Campos del Proceso Base da origen a las relaciones, **Relación (6)** que indica a los Pasos que pertenecen a una Plantilla de Proceso cual es el Paso Original del que se Copio y la **Relación (7)** que indica de que Campo Original se copio el Campo o Actividad de una Plantilla.

A nivel de clases de la capa del Negocio, este modulo esta compuesto por la clase **GestionProcesos** que se encarga de manejar y gestionar los procesos en CMIN 1.0. La clase **Procesos** que representa un Proceso en CMIN el cual contiene una lista organizada de instancias de la clase **Paso** generando la estructura del Proceso en la capa del Negocio, esta clase utiliza la Clase **Proceso**, que es una entidad del negocio (ver Capa de Lógica de Negocio), para el registro de la información en la tabla **PROCESOS**. La clase **Procesos** gestiona los pasos de un proceso. La Clase **Pasos** se apoya también en una clase tiene entidad del Negocio denominada **OrganizacionDePaso** para su registro en la Base de datos. Y finalmente la clase Pasos es la encargada de gestionar sus campos, clase **Campo**. Esta Clase maneja una entidad del negocio clase **CamposDelPaso** para su persistencia. El diagrama de clases de este modulo se pude ver en la Figura 14. Diagrama de Clases del Modulo de Gestión de Procesos.

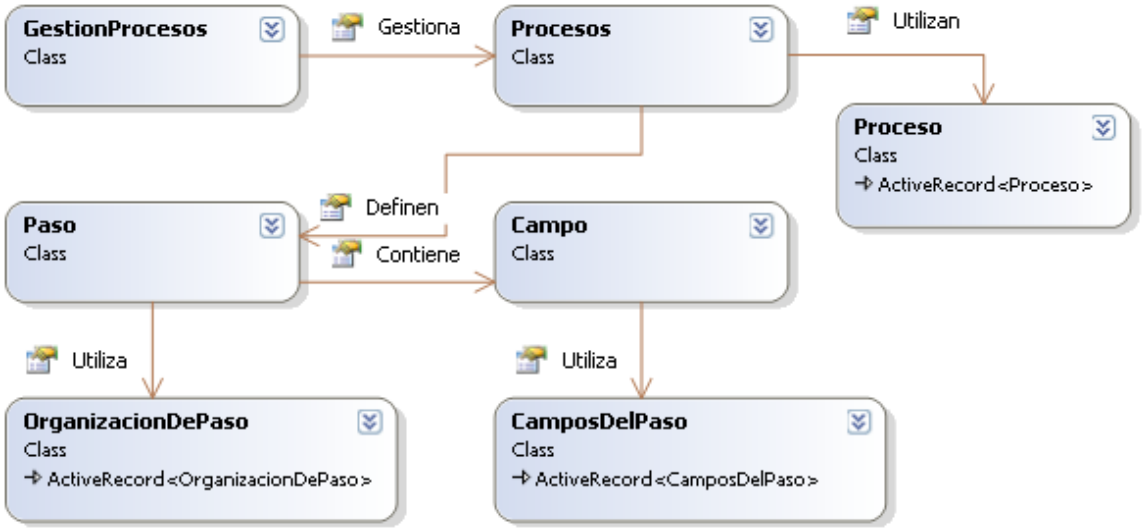

Figura 14. Diagrama de Clases del Modulo de Gestión de Procesos.

9.2 CMIN: Modulo de Gestión de Procesos

En la Figura 15. Edición de Procesos (Información Base) se presenta como el Modulo Permite Editar los Procesos del Sistema y como ejemplo aparece el Ingreso del Proceso CRISP-DM V1.0 (botón **Guardar**), a continuación se presenta la edición de los Pasos del Proceso (botón **Pasos**).

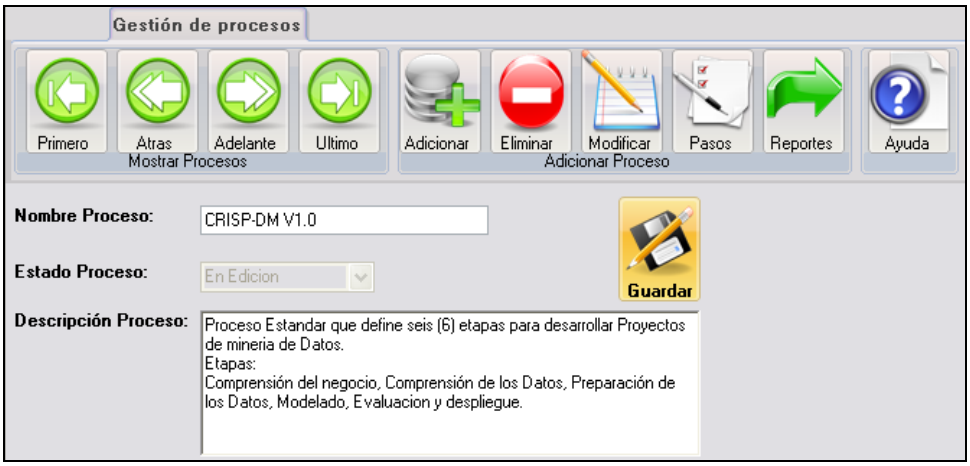

Figura 15. Edición de Procesos (Información Base)

En la Figura 16. Edición de Pasos a partir del Paso Padre. se presenta el Paso Padre tipo metodología que se crea con el Proceso además como este modulo permite adicionar nuevos pasos al Proceso a partir del Paso Padre.

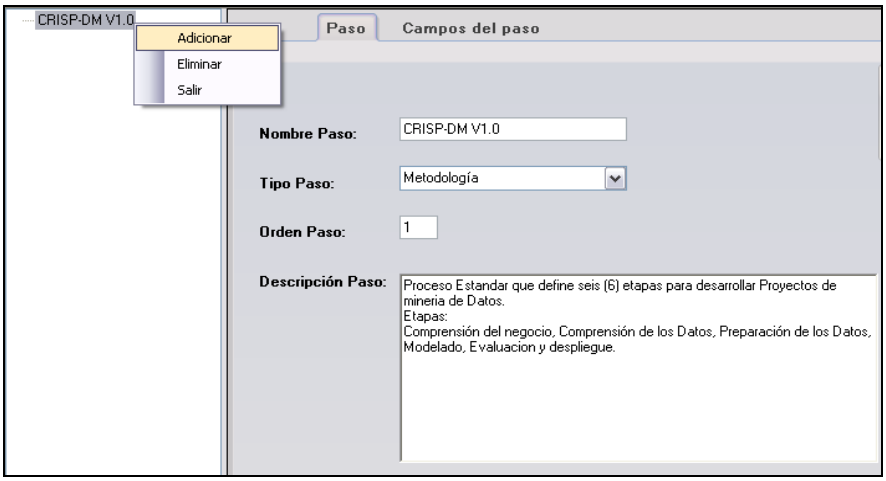

Figura 16. Edición de Pasos a partir del Paso Padre.

La Figura 17. Creando la jerarquía del Proceso presenta como este modulo permite crear pasos hijos (botón **Establecer**) construyendo así la jerarquía.

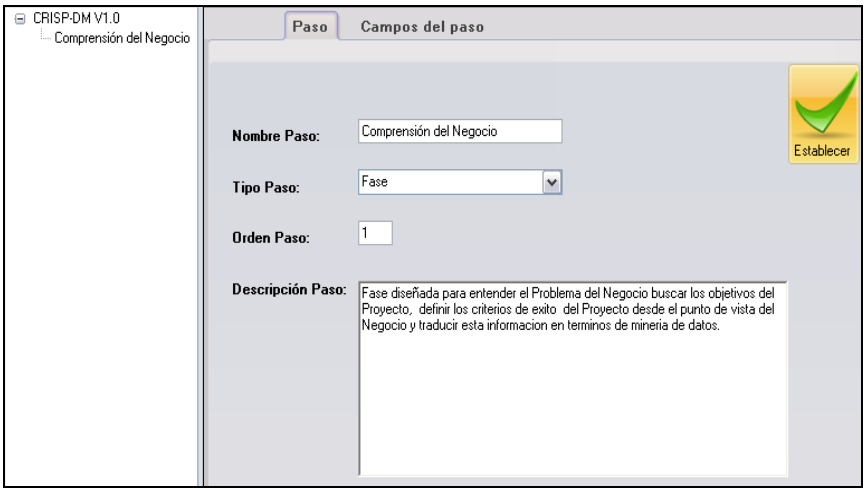

Figura 17. Creando la jerarquía del Proceso.

En la Figura 18. Proceso CRISP-DM 1.0 en CMIN se aprecia el resultado de la Edición de los pasos del Proceso, el Proceso CRISP-DM 1.0 esta incluido en CMIN.

Como se explico anteriormente los Pasos contienen Actividades o Campos los cuales se editan para cada Paso en la pestaña **Campos del Paso**.

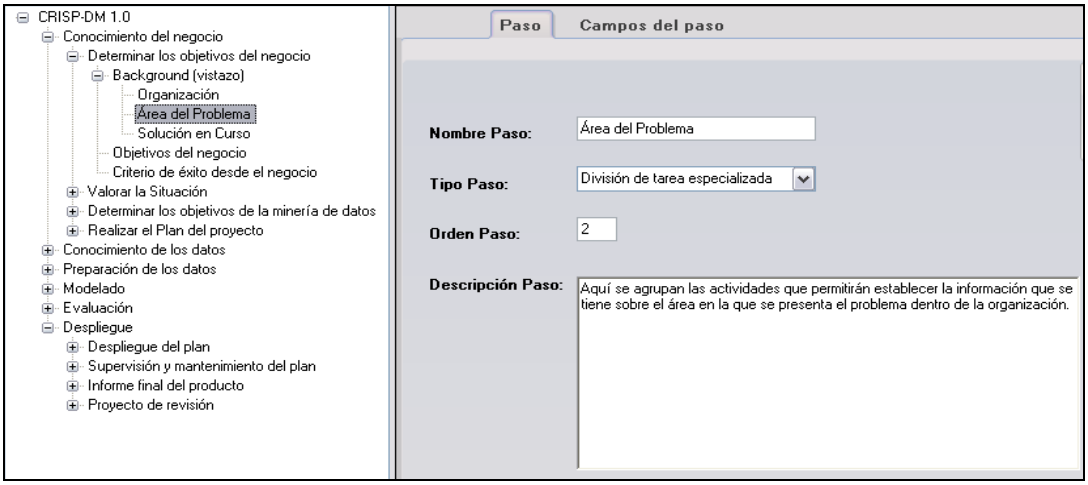

Figura 18. Proceso CRISP-DM 1.0 en CMIN.

En la Figura 19. Edición de Campos del Paso. se aprecia que cuando se adiciona un Campo o actividad a un Paso (botón **Adicionar**) entre la información requerida para el Campo se destaca la **Descripción** que contiene que se debe hacer si es una actividad o la sugerencia que presente la metodología en este paso dos campos especiales (señalados en rojo en la Figura 9) que son:

- **Tipo de Campo**: El cual define si el Campo es una Actividad o sugerencia. Para definir el Campo como actividad se selecciona el tipo Texto ó el tipo Plantilla que indica el Modo en que se va a registrar la información de la Actividad cuando esta sea ejecutada en los Proyectos. Si se selecciona tipo Plantilla el sistema permite

subir un archivo en formato RTF. Archivo en el cual el usuario diseña la mejor forma de registrar la Información de la Actividad.

- **Utiliza Work Flow**: El Modulo Work Flow (Explicado en Detalle en el capitulo CMIN: Work Flow de minería de datos) en CMIN permite al usuario realizar labores de minería de datos. Con la descripción de la actividad el usuario evalúa si para realizar la actividad requiere alguna labor de minería de datos. Por ejemplo cargar datos, Explorar los datos, procesar los datos y crear un modelo son algunas actividades de la Metodología que requieren labores de minería. Si el usuario Chequea esta opción indica que el campo o actividad al ejecutar el Proyecto podrá utilizar el Modulo de Work Flow para su realización.

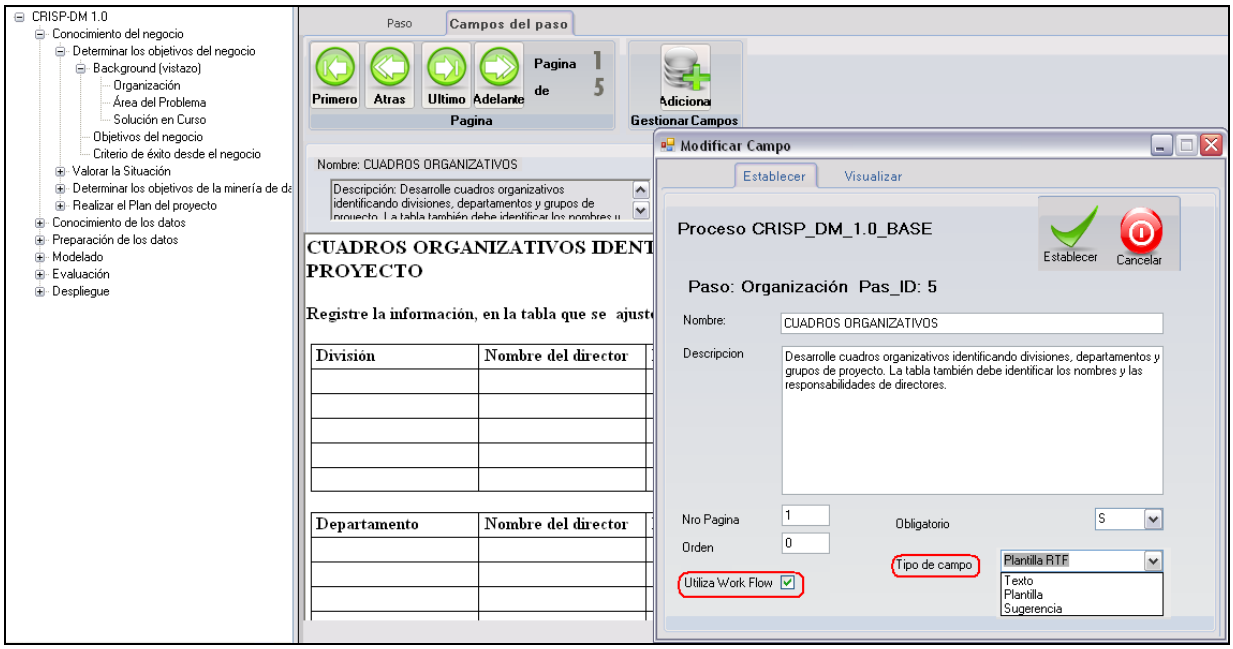

Figura 19. Edición de Campos del Paso.

En la creación de la Plantilla se relaciona el Proceso BASE con el cual el usuario mediante la labor de Mapeo (botón **Pasos**) crea el Proceso específico orientado a un contexto de minería de datos, en este caso el "Marketing" ejemplo planteado en la Figura 20. Edición de Plantillas de Procesos. Este manejo se realiza similar como se hace en los procesos guardando en la tabla **PROCESOS** (ver Figura 15. Edición de Procesos (Información Base)), la diferencia es el tipo de proceso el cual se guardara como plantilla.

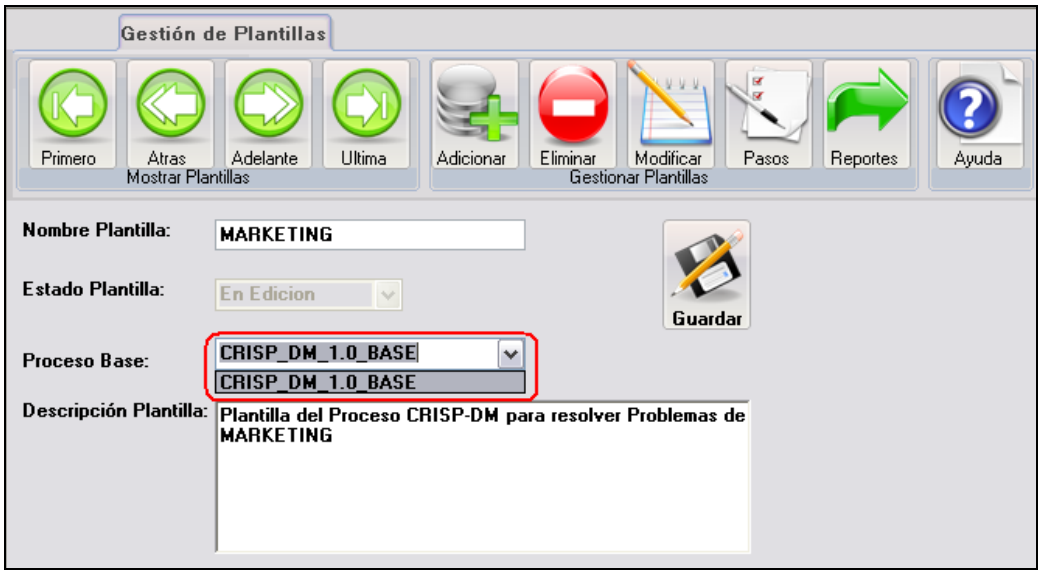

Figura 20. Edición de Plantillas de Procesos

El ejemplo en la Figura 21. Mapeo del Proceso BASE a un Proceso Especifico. presenta la creación de la Plantilla de Proceso MARKETING y como este Modulo permite realizar la labor de Mapeo eliminando los Pasos que para el contexto MARKETING no sean de gran utilidad. Por ejemplo si un usuario o una Empresa es especialista en desarrollar proyectos de un contexto de minería de datos y esta realizando la Plantilla del Proceso Basada en CRISP-DM para ese contexto, y dado un Paso **X** del Proceso BASE cuyo objetivo es definir el contexto de Minería de Datos del Problema, mencionado Paso debiere ser eliminado para hacer mas liviano el Proceso obteniendo como resultado una Plantilla del Proceso BASE con los Pasos que se aplican a ese contexto de minería de Datos en especifico.

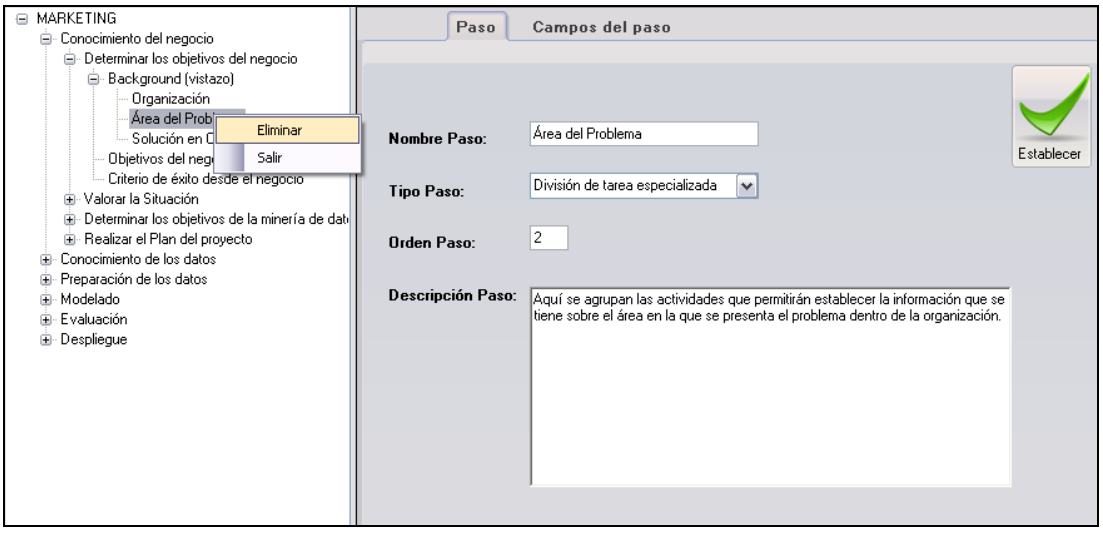

Figura 21. Mapeo del Proceso BASE a un Proceso Especifico.

9.3 Modelo para Modulo de Gestión de Proyectos

Para poder brindar con CMIN 1.0 una herramienta CASE basada en una metodología estándar como es CRISP-DM [14] que soporte y oriente el desarrollo de proyectos de minería de datos, que además permita realizar el seguimiento de un proyecto en cada una de las fases definidas en la instanciación de la metodología seleccionada, así como comprender y ampliar el conocimiento en cuanto al proceso de descubrimiento de conocimiento con Minería de Datos a medida que se desarrolle un proyecto; se creó el modulo de proyectos en CMIN 1.0 para solucionar estos problemas.

Desarrollo de proyectos basados en metodologías

Para lograr que CMIN 1.0 permitiera desarrollar un proyecto de minería de datos basada en una metodología, se decidió hacer que los proyectos heredaran toda la estructura de una metodología ingresada por el modulo de procesos anteriormente explicado, de tal manera que muchos proyectos se pudieran basar de una de las metodologías ingresadas, esto se puede observar en las Figura 22. Relación entre las metodologías o procesos agregados a CMIN 1.0 y los proyectos creados., y la Figura 27. Modulo de gestión de proyectos de la Herramienta CMIN 1.0

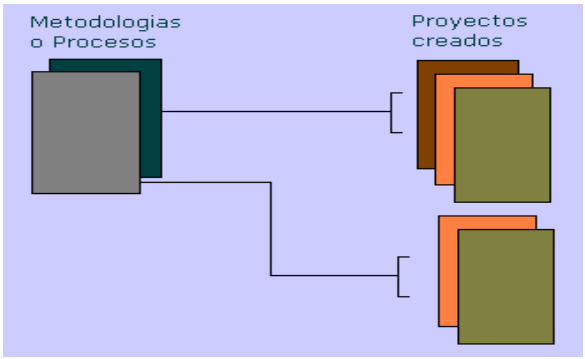

Figura 22. Relación entre las metodologías o procesos agregados a CMIN 1.0 y los proyectos creados.

Para explicar el modelo de base de datos que soporta el desarrollo de Proyectos se utilizaran las etiquetas de las relaciones ("**Relación (1…4)**") que se presentan en la Figura 23. Diseño de la Base de Datos para el manejo de proyectos.

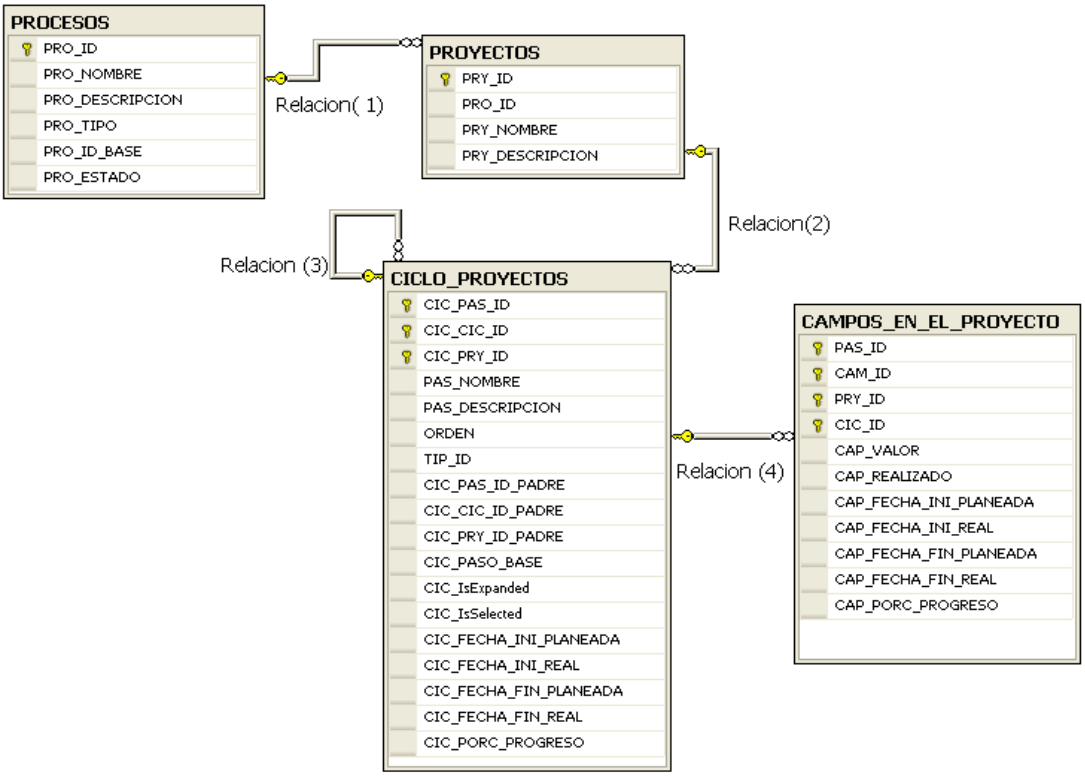

Figura 23. Diseño de la Base de Datos para el manejo de proyectos.

En la tabla **PROYECTOS** se guarda la información básica de un proyecto, la **Relacion (1)** establece que Proceso registrado en la tabla **PROCESOS (**la definición de esta tabla se puede ver Figura 13. Modelo de Base de Datos de Gestión de Procesos.**)** utiliza el proyecto como base para su realización. En la tabla **CICLO_PROYECTOS** se copia toda la estructura de un Proceso o metodología, tomada de la tabla **ORGANIZACION_DE_PASOS** (la definición de esta tabla se puede ver Figura 13. Modelo de Base de Datos de Gestión de Procesos.), esta copia permite que un proyecto posea todos los pasos de una metodología y posibilita la creación de **Ciclos** del Proyecto para un paso, copiando la estructura para el paso en esta misma tabla creando nuevos registros con diferente Identificador pero enlazados tanto al Proyecto como a la jerarquía para su ejecución. La **Relacion (2)** establece que pasos pertenecen a un proyecto específico. La **Relacion (3)** permite mantener la jerarquía planteada ya que es una relación recursiva la cual maneja la relación de padre e hijo entre los pasos y la **Relacion (4)** especifica los campos de cada paso los cuales están en la tabla **CAMPOS_DEL_PROYECTO** en esta tabla se almacena la información resultante de la ejecución de las actividades del proyecto.
La tabla **CICLO_PROYECTOS** contiene copias de las estructuras de los Procesos o metodologías permitiendo el manejo de ciclos en el proyecto sin afectar la estructura original de la metodología. El manejo de **Ciclos** se realiza por medio del campo **CICLO_PROYECTOS.CIC_CIC_ID**. Los Ciclos del Proyecto pueden ser creados para cualquier Paso donde se necesite volver a realizar alguna labor y comparar los resultados de una misma actividad en diferentes Ciclos, o bien realizar de nuevo una actividad o conjunto de actividades que quedaron mal pero no se corrigen para tener un punto de referencia de cómo no se debe realizar la ejecución

de algunas actividades lo cual puede servir para proyectos futuros, esto se puede observar en la Figura 28. Desarrollo de proyectos en la Herramienta CMIN 1.0.

Realizar el seguimiento de un proyecto

El seguimiento de proyectos en CMIN 1.0 se realizo mediante la visualización de diagramas de Gantt, la información necesaria para realizar estos diagramas se obtiene mediante el manejo de las fechas de planeación, realización y el porcentaje de avance de las actividades; las fechas de planeación son las que se definen inicialmente para el desarrollo de una actividad, las fechas de realización son las fechas reales en las que se realizo la actividad, el porcentaje de avance establece el estado de la actividad en cuanto a la planeación.

El manejo de esta información se modeló en la Base de Datos agregando los campos **CAP_FECHA_INI_PLANEADA**, **CAP_FECHA_FIN_PLANEADA**, **CAP_FECHA_INI_REAL**, **CAP_FECHA_FIN_REAL** y **CAP_PORC_PROGRESO** en la tabla **CAMPOS_EN_EL_PROYECTO** la cual guarda la información que es modificada en los campos de los pasos, para manejar la información de los pasos se agregaron los campos **CIC_FECHA_INI_PLANEADA**, **CIC_FECHA_FIN_PLANEADA**, **CIC_FECHA_INI_REAL**, **CIC_FECHA_FIN_REAL** y **CIC_PORC_PROGRESO** en la tabla **CICLO_PROYECTOS** para actualizar la información de los pasos a medida que se actualizan los datos en los campos, esto se puede observar en la Figura 24. Diseño de la Base de Datos para el manejo del seguimiento de un proyecto.

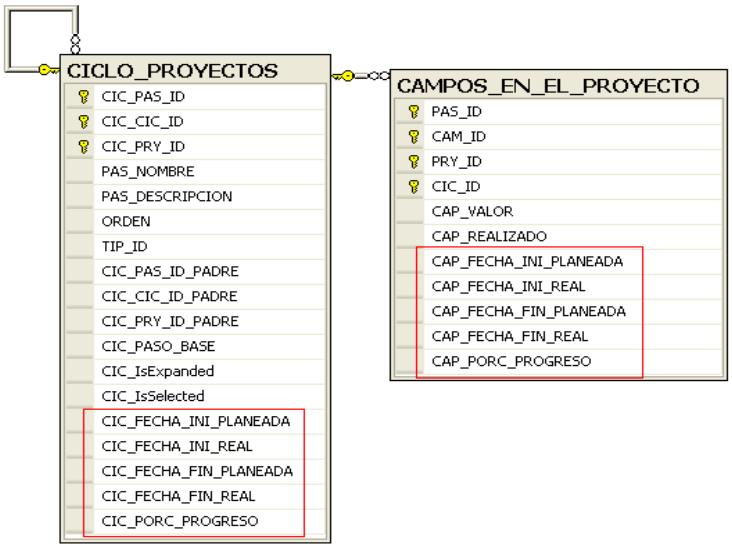

Figura 24. Diseño de la Base de Datos para el manejo del seguimiento de un proyecto.

La modificación tanto de las fechas como del porcentaje de realización se efectúa en los campos de los pasos, ya que es en estos en los que se realizan las actividades de un proyecto, los pasos obtienen la fecha inicial y final de planeación y realización de la menor y mayor fecha de planeación y realización respectivamente de sus pasos hijos o si es el caso de sus campos, el porcentaje de realización de los pasos se calcula de la suma de porcentajes de sus pasos hijos o si es le caso de sus campos,

esto se puede observar en la Figura 29. Modificación de las fechas de planeación, realización y porcentaje de ejecución en la Herramienta CMIN 1.0..

Para la visualización del reporte de Gantt se utilizo el componen CrystalReport que ofrece Visual Studio 2005, para poder generar este reporte el componente solo necesita un dataset tipado, esta información se obtiene de la tabla **CICLO_PROYECTOS** si se solicita el reporte en un paso y se obtiene de la tabla **CAMPOS_EN_EL_PROYECTO** si se solicita el reporte en un campo, esto se puede observar en la Figura 30. Visualización de los reportes de Gantt en la Herramienta CMIN 1.0..

Comprender y ampliar el conocimiento en cuanto al proceso de descubrimiento de conocimiento

Para apoyar al usuario en su proceso de comprensión o ampliación del conocimiento con respecto al proceso de descubrimiento de conocimiento, se agrego una explicación de cada uno de los pasos de la metodología la cual se le presenta al usuario a medida que la recorre durante la ejecución de un proyecto.

El manejo de esta información se modelo en la Base de Datos agregando el campo **CAM_DESCRIPCION** el cual contiene la descripción del paso según la Metodología CRISP-DM [14], esto se puede observar en al Figura 25. Diseño de la Base de Datos para el manejo de las descripciones de los pasos.

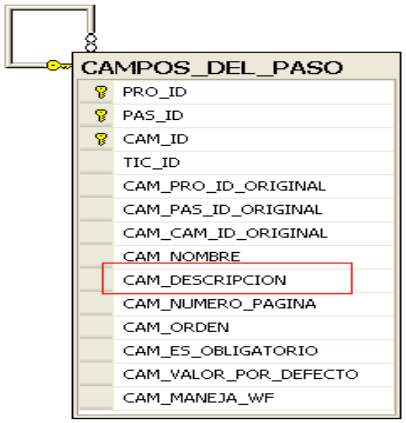

Figura 25. Diseño de la Base de Datos para el manejo de las descripciones de los pasos.

Estas descripciones son muy importantes ya que además de permitir al usuario saber que esta realizando, le permite conocer y entender un poco mas de la metodología que esta utilizando (ver Figura 31. Descripción de los pasos de una metodología en la Herramienta CMIN 1.0.), haciendo que en futuros proyectos la calidad sea mejor y se puedan conseguir mejores resultados en estos.

Adicionalmente hay campos que permiten la ejecución del modulo de WorkFlow (WF) ofrecido por CMIN 1.0 el cual permite que lo usuarios puedan realizar practicas propias de minería de datos, este modulo se tratara mas adelante.

A nivel de clases de la capa del Negocio, el modulo de gestión de proyectos contiene la clase **GestionProyectos** que se encarga de manejar y gestionar los proyectos en CMIN 1.0. La clase **Proyectos** que realiza funcionalidades sobre los proyectos como

creación de ciclos lo que conlleva a crear nuevos Pasos en el Proyecto. Esta clase contiene una instancia de la clase **Proyecto** que es una entidad del negocio (ver Capa de Lógica de Negocio), la cual permite la modificación de la tabla **PROYECTOS**. Para gestionar los pasos de un proyecto se creo la clase **PasoProyecto**, que representa un paso en el proceso base del proyecto, esta clase utiliza las clases entidad del Negocio denominadas **CicloProyecto** y **CamposDelPaso** para actualización y cargue de la información de las tablas **CICLO_PROYECTOS** y **CAMPOS_DEL_PASO.** Finalmente la Clase **PasoProyecto** gestiona los campos o actividades del Proyecto, la Clase **CampoProyecto** mediante la Entidad del Negocio **CamposEnElProyecto** actualiza la tabla **CAMPOS_EN_EL_PROYECTO**. El diagrama de clases de este modulo se pude ver en la Figura 26. Diagrama de Clases del Modulo de Gestión de Proyectos.

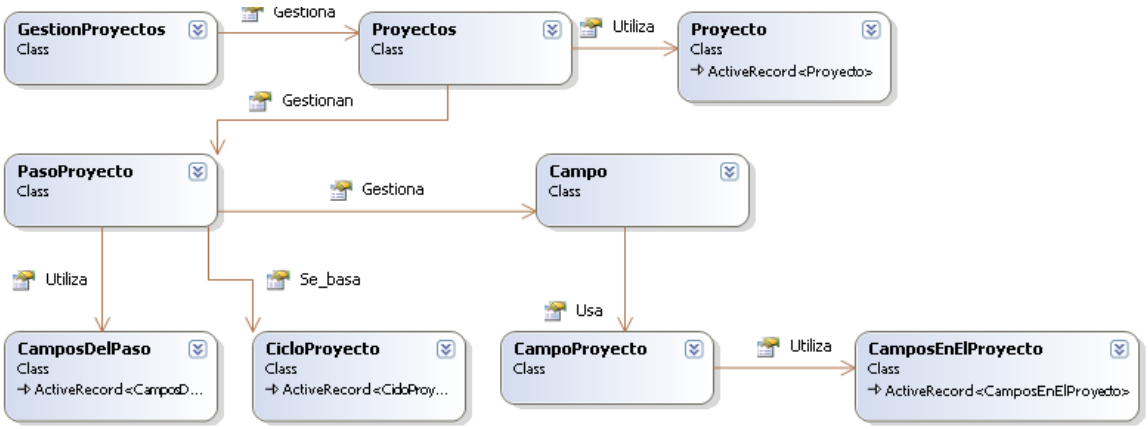

Figura 26. Diagrama de Clases del Modulo de Gestión de Proyectos.

9.4 CMIN: Modulo Gestión de Proyectos

En la Figura 27. Modulo de gestión de proyectos de la Herramienta CMIN 1.0**.** se presenta como este Modulo Permite crear los Proyectos y como ejemplo aparece la creación del Proyecto "Proyecto de Análisis" (botón **Guardar**), a continuación se puede comenzar el desarrollo del proyecto (botón **Pasos**).

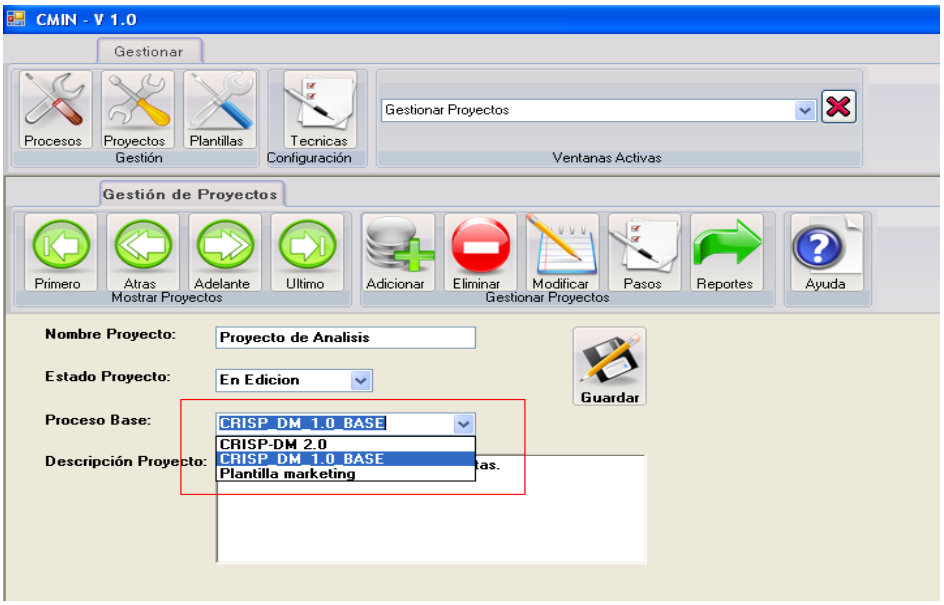

Figura 27. Modulo de gestión de proyectos de la Herramienta CMIN 1.0.

En la Figura 28. Desarrollo de proyectos en la Herramienta CMIN 1.0. se presenta el desarrollo de un proyecto, en el numeral **(1)** se puede observar la estructura de la metodología la cual es usuario la recorre a medida que esta desarrollando un proyecto de minería en CMIN 1.0, en el numeral **(2)** se observa la sección donde a parecen los campos a desarrollar pertenecientes al paso en el cual se encuentren, el numeral **(3)** se encuentra el botón que guarda la información del campo modificado, el numeral **(4)** se observa como se puede crear un ciclo en un proyecto, en el numeral **(5)** vemos como se visualizan los ciclos creados.

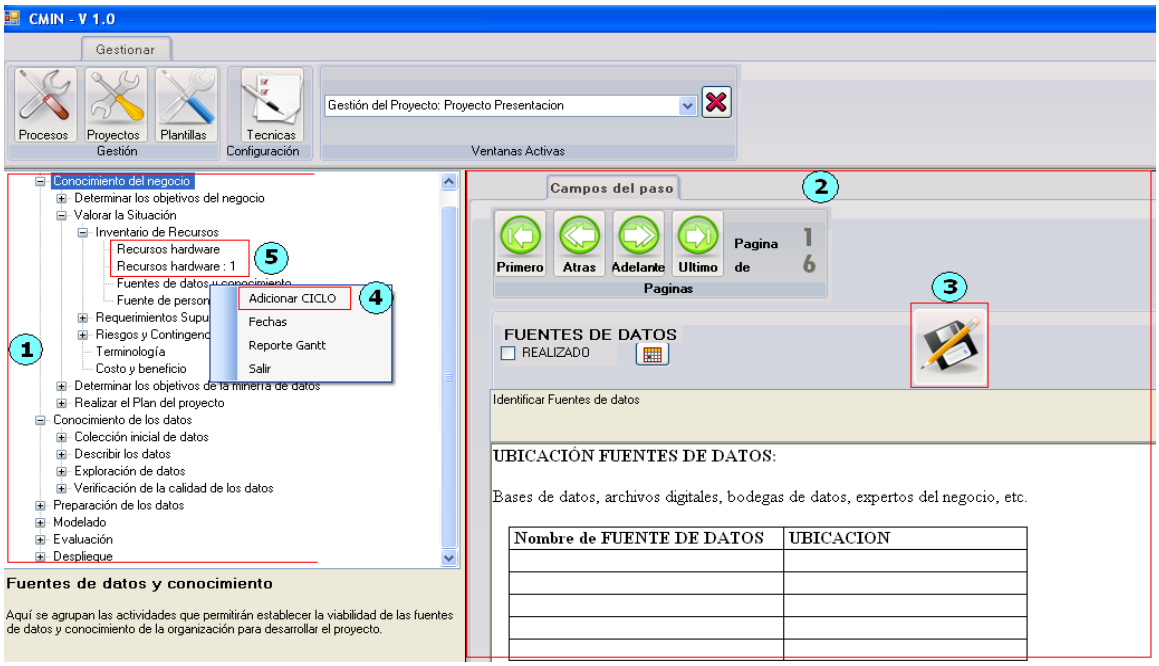

Figura 28. Desarrollo de proyectos en la Herramienta CMIN 1.0.

En la Figura 29. Modificación de las fechas de planeación, realización y porcentaje de ejecución en la Herramienta CMIN 1.0. se presenta la forma en la que se puede realizar seguimiento a un proyecto en desarrollo, en el numeral **(1)** se puede observar el botón que ejecuta la opción que permite modificar el porcentaje de realización, así como las fechas de planeación y realización, en el numeral **(2)** se observa la sección donde se encuentran las fechas para modificarlas, el numeral **(3)** se observa la sección en la cual se puede modificar el porcentaje de realización.

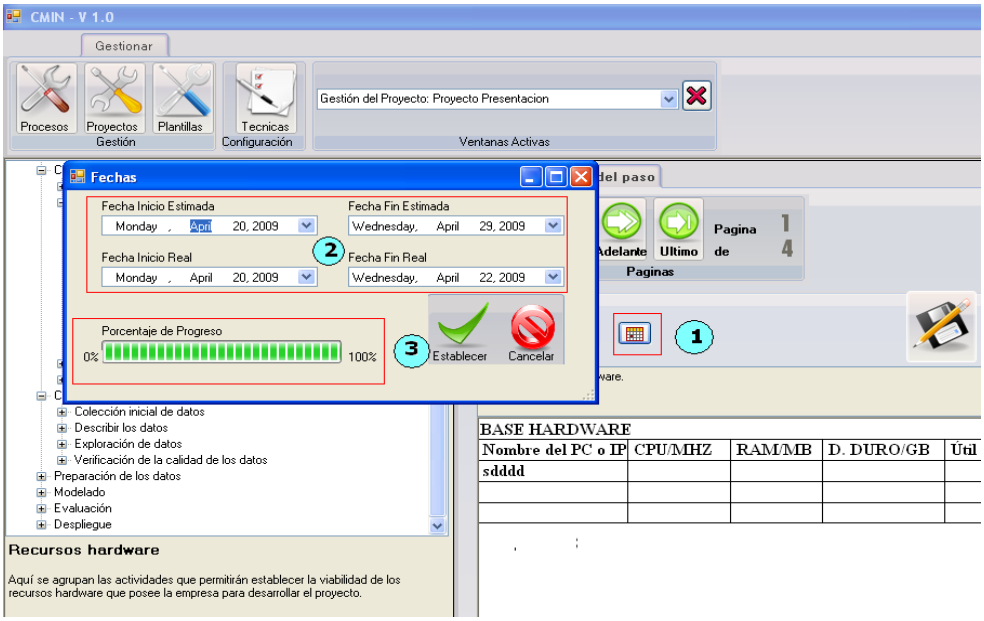

Figura 29. Modificación de las fechas de planeación, realización y porcentaje de ejecución en la Herramienta CMIN 1.0.

En la Figura 30. Visualización de los reportes de Gantt en la Herramienta CMIN 1.0. se presenta la forma en la que se puede visualizar el reporte de Gantt para realizar en análisis del seguimiento del proyecto, en el numeral **(1)** se puede observar el menú que me permite ejecutar la opción que permite visualizar le reporte, en el numeral **(2)** se observa visualización del reporte.

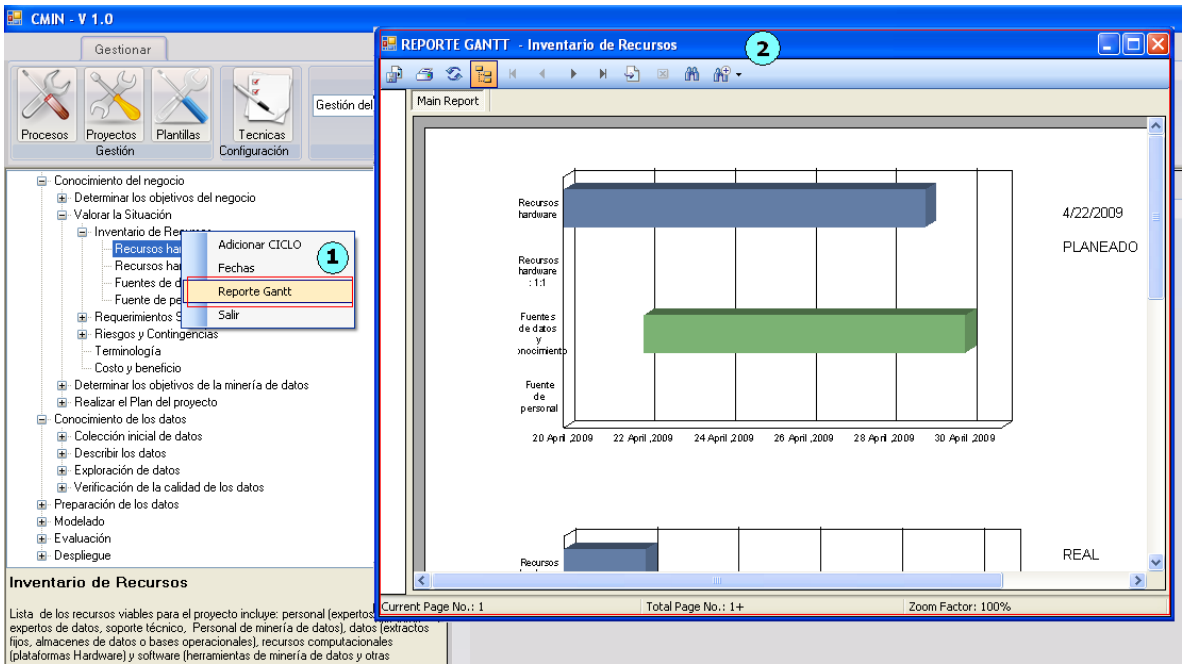

Figura 30. Visualización de los reportes de Gantt en la Herramienta CMIN 1.0.

En la Figura 31. Descripción de los pasos de una metodología en la Herramienta CMIN 1.0. se presenta la forma en la que se puede visualizar la descripción de los pasos y los campos de tal manera que el usuario sepa lo que esta haciendo en ese momento, en el numeral **(1)** se puede observar la descripción que se presenta para los pasos, en el numeral **(2)** la descripción que se presenta para los campos, en el numeral **(3)** se observa el botón que ejecuta la opción para poder trabajar en el WorkFlow.

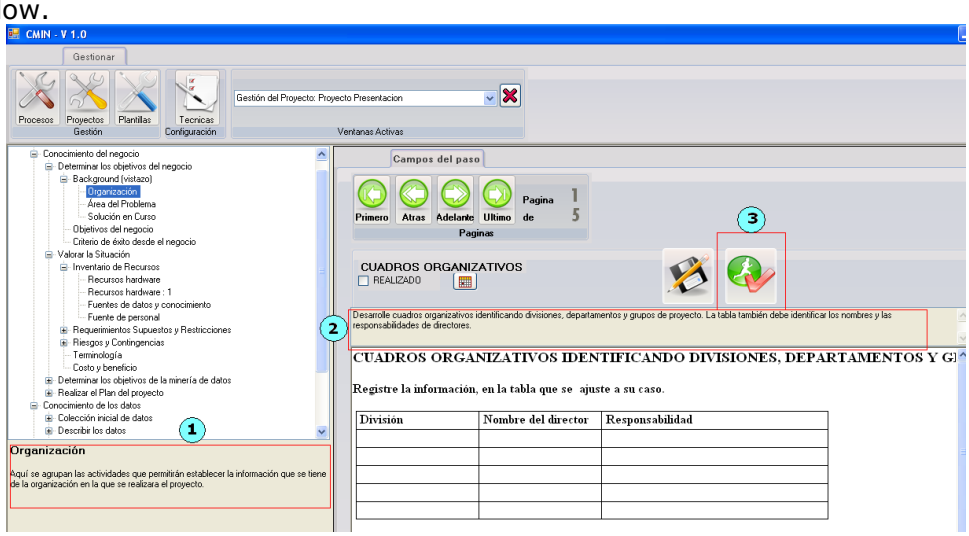

Figura 31. Descripción de los pasos de una metodología en la Herramienta CMIN 1.0.

10 CMIN: Work Flow de minería de datos

La Herramienta CASE CMIN tiene varias características especiales, entre las que se encuentra la extensibilidad del conjunto de algoritmos que ofrece en su WORK FLOW

de minería de datos. Este WORK FLOW es utilizado por los usuarios para realizar las labores de minería de datos en las Actividades de los Proyectos que así lo requieran. El Modulo de WORK FLOW en CMIN soporta el proceso de adición de algoritmos y la utilización de los mismos en el WORK FLOW.

10.1 Modulo de WORK FLOW

Para explicar como funciona este modulo se describirá en primera instancia la interfaz grafica del WORK FLOW, seguido del Modelo de Base de Datos y a partir de esta descripción se centrará en los siguientes temas:

- Tipos de Objetos del WORK FLOW y las relaciones entre ellos.
- Proceso de adición de un algoritmo a un Tipo de Objeto del WORK FLOW de CMIN
- Invocación de Métodos de Algoritmos en CMIN.
- Relación entre el WORK FLOW y las actividades de los Proyectos.

En la Figura 32. WORK FLOW de minería en CMIN el numero (1) señala los **tipos de objetos** del WORK FLOW, el numero (2) un **objeto ofrecido** del tipo de objeto "Fuente de Datos" y el numero (3) el **objeto en ejecución** del WORK FLOW el cual es configurado y utilizado por el usuario ver Figura 32. WORK FLOW de minería en CMIN.

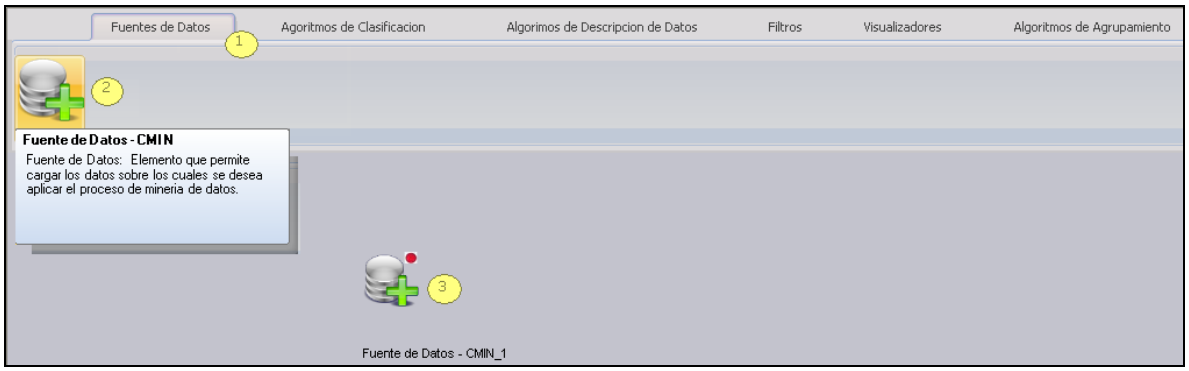

Figura 32. WORK FLOW de minería en CMIN.

Este modulo esta soportado en el esquema de Base de datos que se muestra en la Figura 33. Modelo de Base de Datos del Modulo de WORK FLOW.

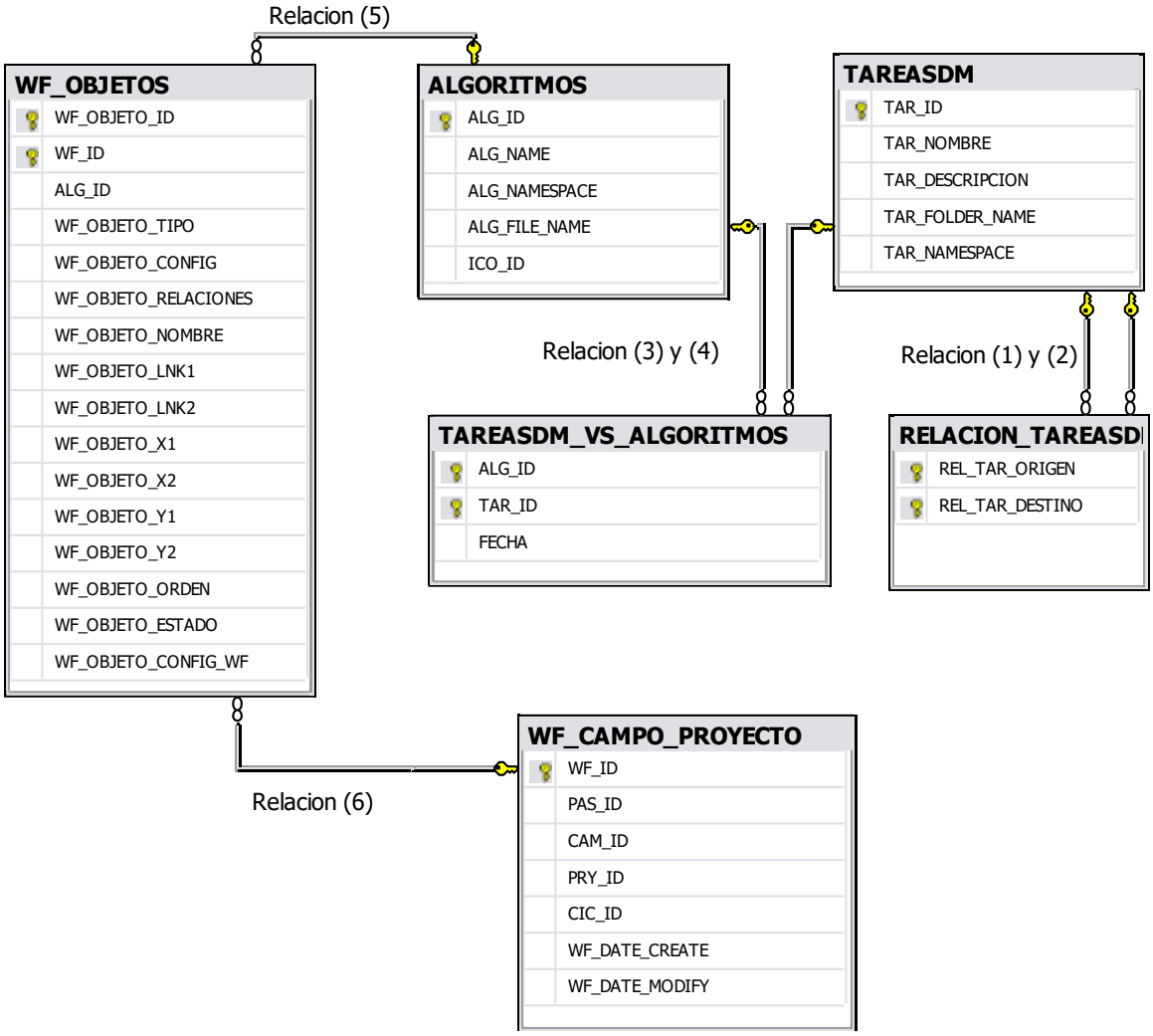

Figura 33. Modelo de Base de Datos del Modulo de WORK FLOW.

Los tipos de Objetos del WORK FLOW son almacenados en la Tabla **TAREASDM**, las relaciones **(1)** y **(2)** indican los tipos de objetos que se pueden enlazar entre si, registrando en la Tabla **RELACION_TAREASDM** el identificador del tipo de objeto destino y el del tipo de objeto origen del enlace, cada registro en esta tabla es un enlace entre los tipos de objetos. Los algoritmos se registran en la tabla **ALGORITMOS** y las relaciones **(3)** y **(4)** definen a que Tipos de Objeto del WORK FLOW pertenecen, registrando en la tabla **TAREASDM_VS_ALGORITMOS** esa relación, de esta manera los algoritmos son presentados en el WORK FLOW como **objetos ofrecidos** (ver Figura 32) de los tipos de objetos. El WORK FLOW es utilizado por los usuarios en algunas actividades del Proyecto para lo cual adicionan los **objetos ofrecidos** a la zona de trabajo, convirtiéndolos en **objetos en ejecución** (ver Figura 32). Estos objetos en ejecución son almacenados en la tabla **WF_OBJETOS** y la **Relación (5)** indica que Algoritmo utiliza este objeto para su ejecución. Los objetos en ejecución pertenecen a un WORK FLOW desarrollado por el usuario para una actividad del Proyecto el cual es registrado en la tabla **WF_CAMPO_PROYECTO** la **Relación (6)** Indica que objetos en ejecución pertenecen al WORK FLOW de la actividad.

10.1.1Tipos de Objetos del WORK FLOW y relaciones entre ellos

Los tipos de objetos en CMIN están definidos en dos grupos los basados en los tipos de Problemas de minería de datos (ver Tipos de Problemas de Minería de Datos) y los basados en las herramientas necesarias para las etapas de comprensión y procesamiento de datos. Obteniendo el siguiente listado:

Tipos de Problemas de minería de datos
- Algoritmos de Clasificación
- Algoritmos de Descripción de Datos
- Algoritmos de Regresión Lineal
- Algoritmos de Agrupamiento

Fases Comprensión y Procesamiento de Datos
- Fuentes de Datos
- Filtros
- Visualizadores

Con el objetivo de permitir adicionar algoritmos u objetos a los tipos de objetos en tiempo de ejecución, para cada tipo de objeto del WORK FLOW se le definió una Interfaz de Software (ver Interfaces de Software (Interface)), que agrupe los métodos necesarios para su aplicación y otros métodos de Interacción con los demás Tipos del WORK FLOW.

En tiempo de desarrollo para registrar (ver Figura 34. Edición de Tipos de objetos del WORK FLOW) un tipo de objeto del WORK FLOW en CMIN es necesario que la Interface sea convertida a un Assembly (ver Figura 35. Generar Assembly) en nuestro caso se realizo mediante un Proyecto de Librería de Visual Studio [42].

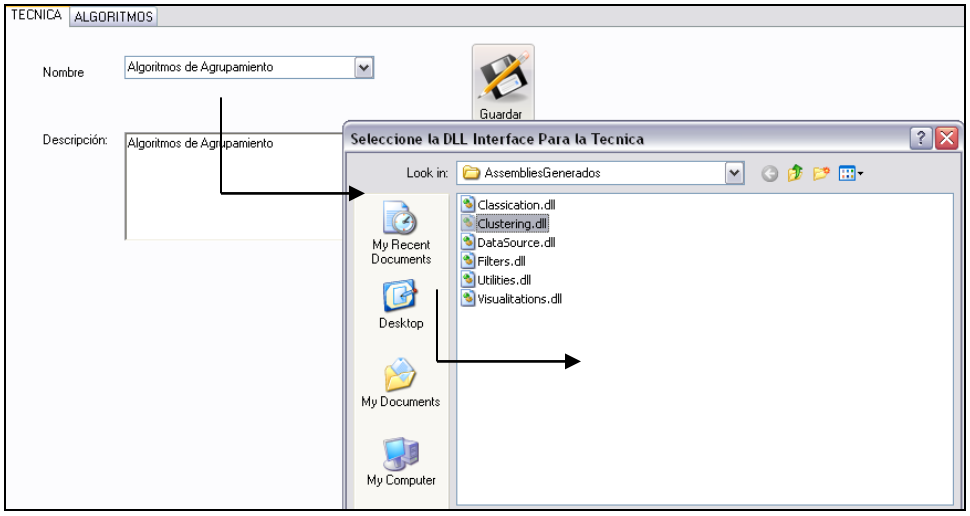

Figura 34. Edición de Tipos de objetos del WORK FLOW.

Figura 35. Generar Assembly.

La información del tipo de objeto es almacenado en la Base de Datos y el archivo ".dll" es copiado y almacenado en la carpeta local de CMIN **Assemblies_CMIN**.

Después de Ingresar el tipo de objeto se le debe definir con quien puede establecer enlaces, es decir definir qué tipo de objeto puede entregarle información y a que tipo de objeto le puede entregar información (ver Figura 36).

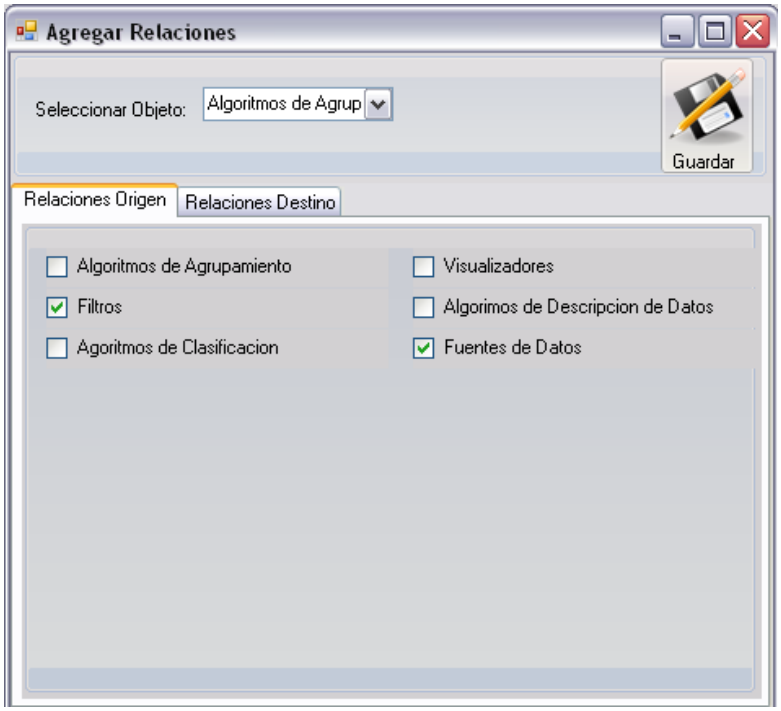

Figura 36. Edición de Relaciones de Tipos de Objetos del Work Flow de CMIN.

10.1.2 Proceso de adición de un algoritmo a un Tipo de Objeto del WORK FLOW de CMIN

El proceso para adicionar un algoritmo o nuevo **objeto ofrecido** a un **tipo de objeto** de CMIN es el siguiente:

- El usuario Programador crea un Proyecto de Librería en Visual Studio en cual debe adicionar como referencia la dll que se generó a partir de la interfaz de Software creada para el Tipo de objeto, para el cual se va a crear el Nuevo objeto ver Figura 37. Generar Assembly nuevo Algoritmo.

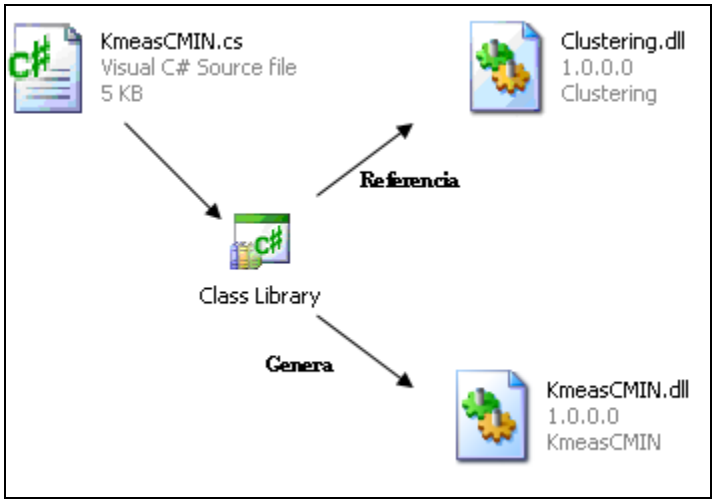

Figura 37. Generar Assembly nuevo Algoritmo.

- El usuario programa su algoritmo en el Proyecto de Librería, implementando la Interfaz definida por CMIN para el tipo de objeto al cual va a ser adicionado (ver Figura 38. Diagrama de Clases en el Proyecto de Librería.), generando una **nueva dll**, que contiene la funcionalidad del nuevo Algoritmo para CMIN. Las dll resultado de este proyecto son comprimidas en un .zip ver Figura 39. dll resultado del Proyecto de librería.

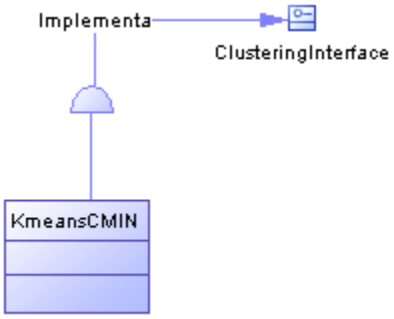

Figura 38. Diagrama de Clases en el Proyecto de Librería.

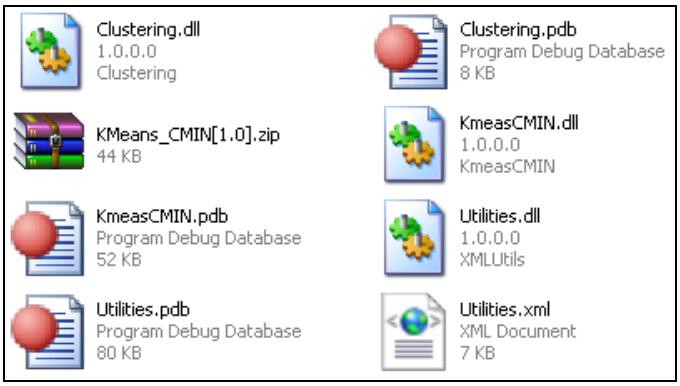

Figura 39. dll resultado del Proyecto de librería.

- El usuario que desee agregar este nuevo algoritmo a CMIN utiliza este modulo, el cual solicita el .zip que contiene la dll del algoritmo ver Figura 40. Adición de algoritmo al tipo de objeto Algoritmos de agrupamiento en CMIN..

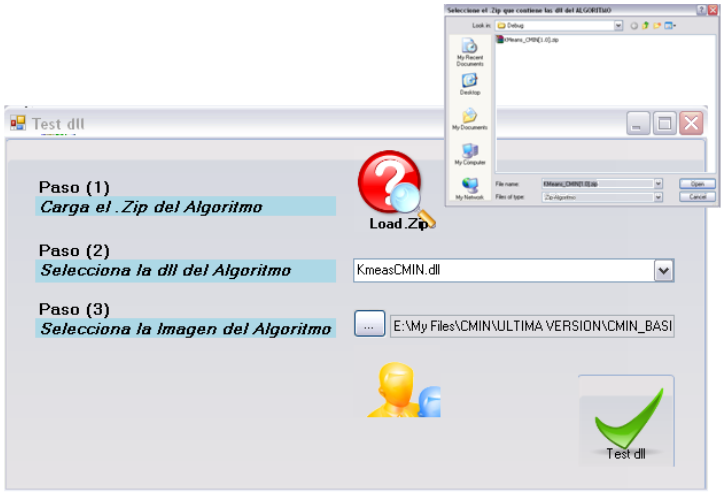

Figura 40. Adición de algoritmo al tipo de objeto Algoritmos de agrupamiento en CMIN.

- El modulo de WORK FLOW valida que ese nuevo algoritmo cumpla con la interfaz Software definida para el Tipo de objeto al que se dese ingresar. Utilizando la dll del tipo de objeto que tiene almacenada en la carpeta local de CMIN **Assemblies_CMIN** y la dll que viene en el .zip especificado, esta comparación se realiza utilizando reflexión (ver Reflexión (System.Reflection)) cargando los Assemblies y comparando los métodos de forma similar a lo presentado en las Tablas Tabla 5 y Tabla 6 con reflexión ver Figura 41. Proceso de validación de algoritmo en CMIN.

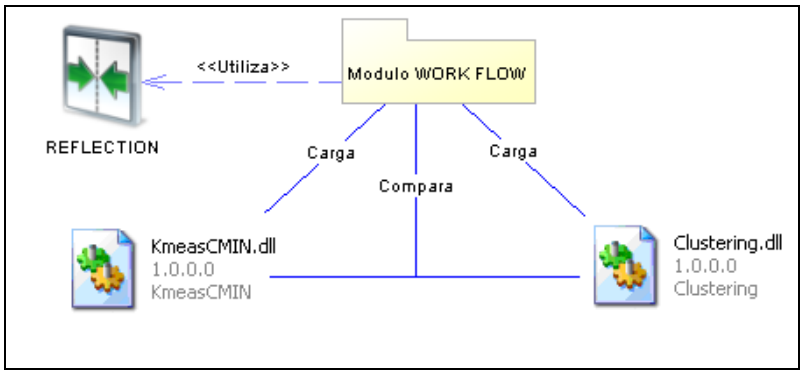

Figura 41. Proceso de validación de algoritmo en CMIN.

- Si el nuevo algoritmo cumple con la Interfaz Software del Tipo de Objeto se registra en la Base de datos y los archivos del .zip son descomprimidos y almacenados en la carpeta local de CMIN **ALGORITMOS**, quedando listo para ser utilizado en el WORK FLOW de CMIN ver Figura 42. Algoritmo Nuevo listo para su utilización.

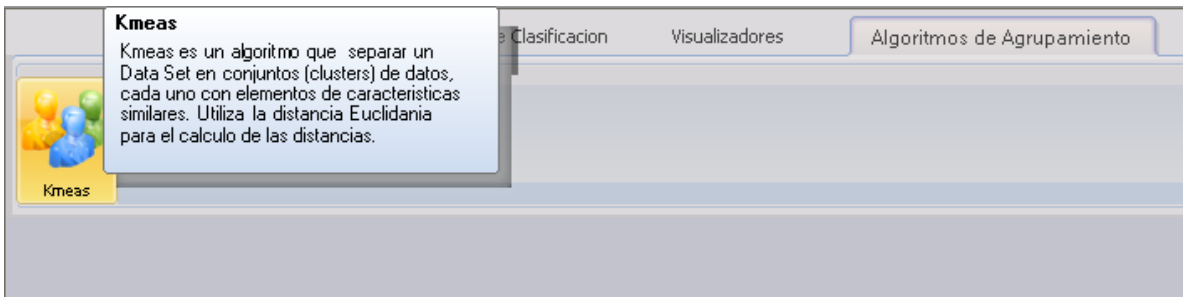

Figura 42. Algoritmo Nuevo listo para su utilización.

10.1.3Invocación de Métodos de Algoritmos en CMIN

Para la invocación de los métodos de las dll de los algoritmos, se debe tener en cuenta que CMIN almacena los Assembly o dll de los algoritmos en carpetas locales y que tiene también almacenados los Asemmblies de los tipos de Objetos. Estos tipos de objetos del WORK FLOW son estáticos y la parte dinámica son algoritmos u objetos de cada uno, los cuales pueden crecer en tiempo de ejecución.

Con este precedente y destacando que nosotros como desarrolladores de CMIN creamos las interfaces de Software de cada tipo de Objeto, teniendo en cuenta métodos que permitieren la interacción de los algoritmos con el usuario y el Núcleo CMIN. Desarrollamos la programación del WORK FLOW basados en la información de las definiciones planteadas en las Interfaces Software (es decir programamos para los tipos de objetos) para la invocación de los métodos de los algoritmos de cada tipo de objeto, utilizando para esto REFLECTION para cargar los Assembly o dll de cada algoritmo, y realizar la invocación de los métodos donde fuere necesario. Programando además la interacción de los objetos basados en las reglas o relaciones definidas en la Figura 36. Edición de Relaciones de Tipos de Objetos del Work Flow de CMIN.

A continuación se presenta un ejemplo de cómo se implemento en CMIN la configuración de los tipos de objeto Fuentes de Datos. La interfaz Software del tipo

de objeto "Fuentes de Datos" contiene la definición del método "configuration" (ver Tabla 8).

```
public interface DataSourceInterface
 {
  .
  Utilities.ParseParameters Configuration();
  .
  .
  bool SetConfiguration(Utilities.ParseParameters Configuration);
  .
  .
  .
 }
```

Tabla 8. Definición del método Configuración en DataSourceInterface.

Para el ejemplo se plante el escenario en el cual se ha adicionado un objeto o algoritmo al tipo de objeto fuente de datos y dicho objeto ha sido adicionado a la zona de trabajo del WORK FLOW convirtiéndolo en un objeto en ejecución y esta seleccionado (ver Figura 43. Escenario del ejemplo).

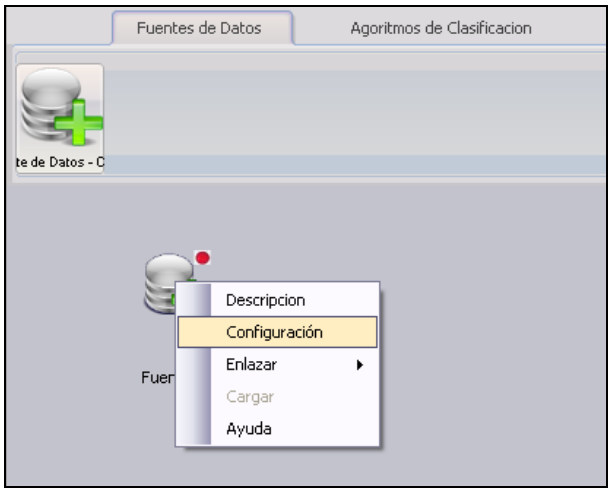

Figura 43. Escenario del ejemplo.

El objeto en ejecución conoce que tipo de objeto es y contiene como propiedad el algoritmo que representa la dll. Cuando el usuario hace clic en Configuración el objeto en ejecución del WORK FLOW le dice al algoritmo que método debe ejecutar, en este caso el método "configuration". A continuación lo que realiza el Algoritmo es mediante REFLECTION carga el Assembly y el TYPE con el NamesPace definido en Figura 40. Adición de algoritmo al tipo de objeto Algoritmos de agrupamiento en CMIN. e invoca el método del algoritmo. La información de retorno que es la configuración del algoritmo para ese objeto, es almacenada en el Objeto en ejecución lo que permite su persistencia y su futura asignación al algoritmo mediante el método "SetConfiguration" definido también en la interfaz del tipo de objeto "Fuente de Datos" (ver Tabla 8).

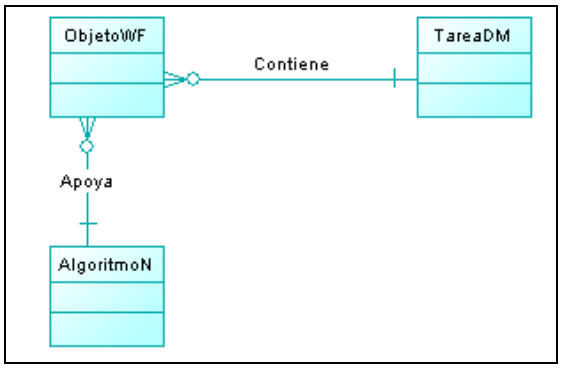

Figura 44. Diagrama de Clases de Ejecución de objetos del WORK FLOW.

10.1.4 WORK FLOW y Actividades de los Proyectos

Como se ha mencionado el WORK FLOW de minería de datos de CMIN puede ser utilizado en las actividades del proyecto. Esta relación se soporta con el esquema presentado en la Figura 45. Relación entre campos del Proyecto y los Work Flows. donde se muestra que cada campo o actividad del Proyecto es posible que contenga un WORK FLOW.

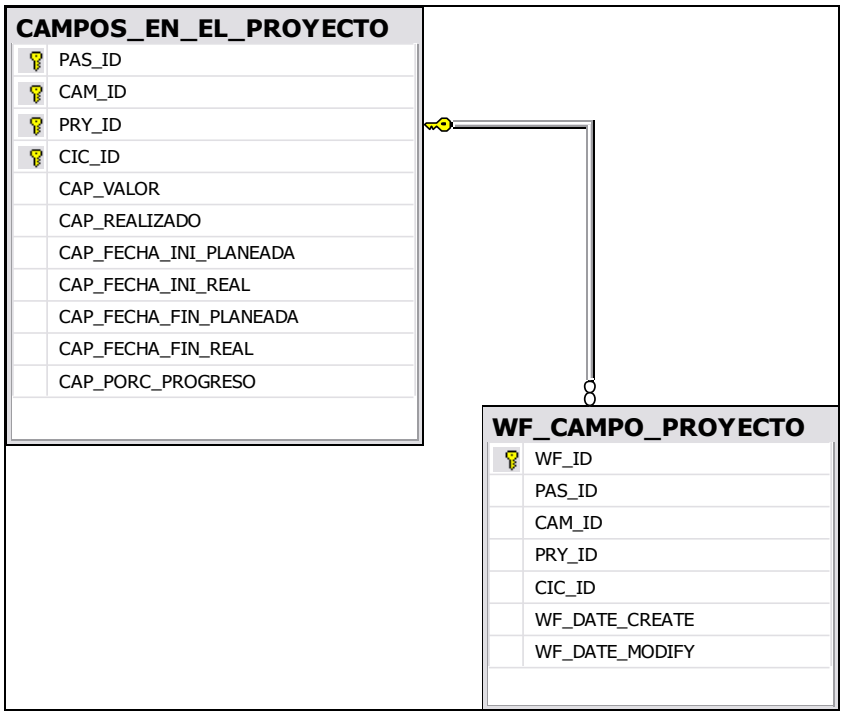

Figura 45. Relación entre campos del Proyecto y los Work Flows.

Este esquema permite que los usuarios desarrollen en cada paso solo la labor planteada en el WORK FLOW. Para un paso siguiente el WORK FLOW estará en blanco pero mediante la utilidad ofrecida en este Modulo para copiar el WORK FLOW de las actividades del Proyecto a la actividad actual, se puede recuperar lo trabajado, adicionar lo que la actividad propone, dejando así un rastro del Proceso que se siguió para llegar hasta el Modelo final mejorando el Proceso de revisión en cada caso y su entendimiento.

Para citar un ejemplo la actividad Como Extraer los Datos, puede manejar el WORK FLOW como se puede ver en la Figura 46. Actividad COMO EXTRAER LOS DATOS en el Proyecto.

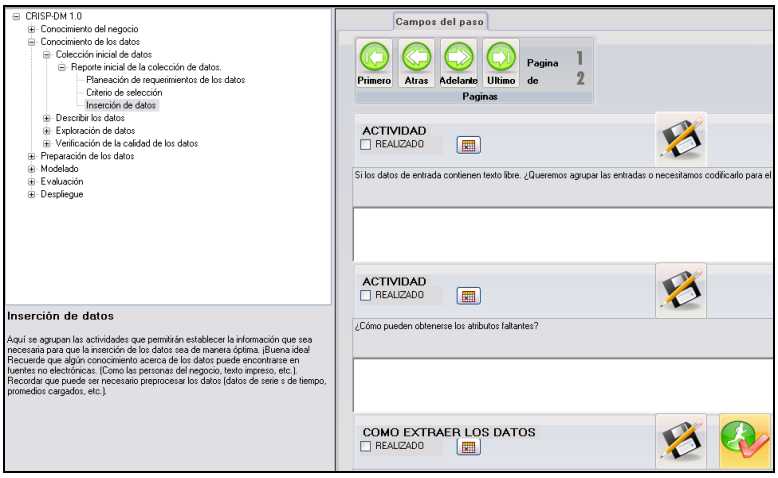

Figura 46. Actividad COMO EXTRAER LOS DATOS en el Proyecto.

Después de ejecutar el WORK FLOW de la actividad Como Extraer los Datos, se puede observar en la Figura 47. WORK FLOW de la Actividad COMO EXTRAER DATOS el resultado de esta ejecución.

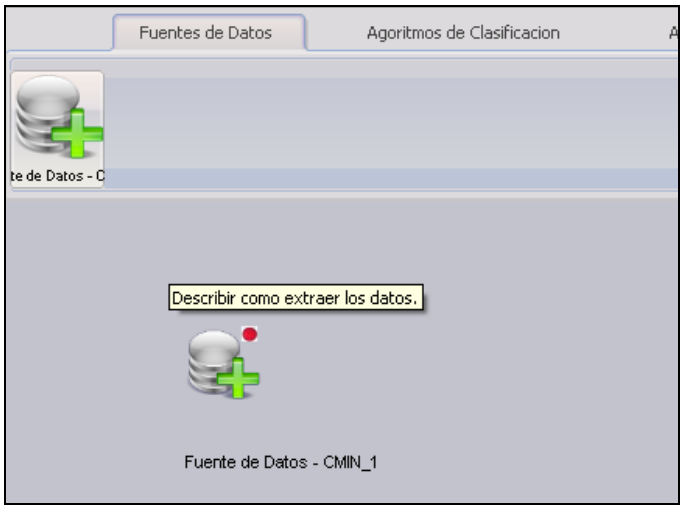

Figura 47. WORK FLOW de la Actividad COMO EXTRAER DATOS.

La Actividad Pasos para transformaciones necesita también utilizar el WORK FLOW ver Figura 48. Actividad PASOS PARA TRANSFORMACIONES en el Proyecto

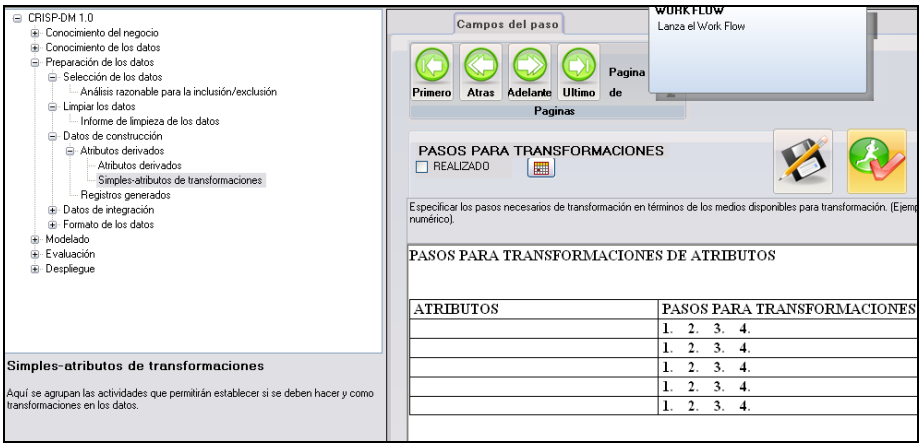

Figura 48. Actividad PASOS PARA TRANSFORMACIONES en el Proyecto.

Originalmente su WORK FLOW esta vacío pero mediante la Utilidad "Copiar Work Flow" puede obtener el WORK FLOW de la actividad COMO EXTRAER LOS DATOS ver Figura 49. WORK FLOW de actividad PASOS PARA TRANSFORMACIONES, Figura 50. Utilidad Copia del WORK FLOW y Figura 51. WORK FLOW de actividad PASOS PARA TRANSFORMACIONES después del Proceso de copia.

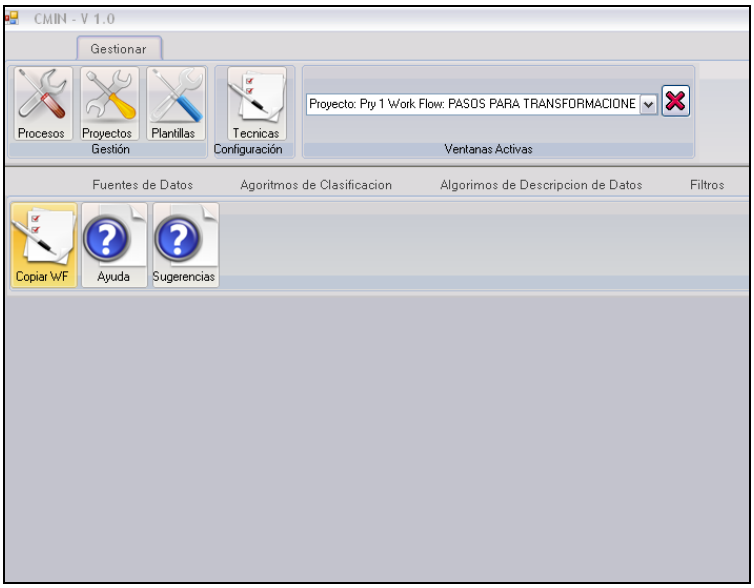

Figura 49. WORK FLOW de actividad PASOS PARA TRANSFORMACIONES.

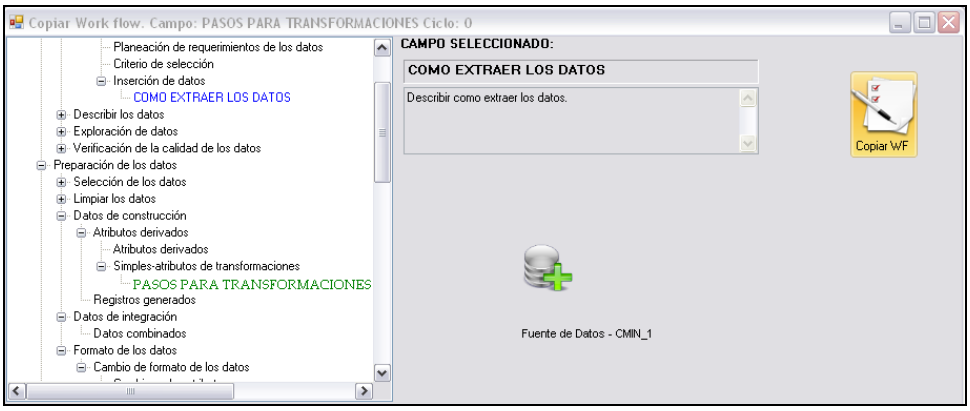

Figura 50. Utilidad Copia del WORK FLOW.

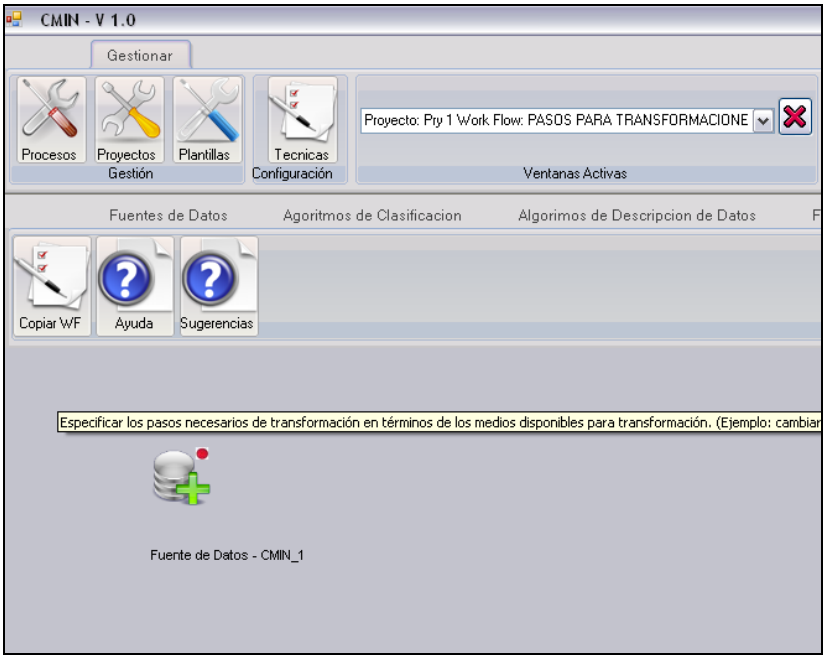

Figura 51. WORK FLOW de actividad PASOS PARA TRANSFORMACIONES después del Proceso de copia.

La copia del WORK FLOW mantiene las configuraciones de los objetos, sus relaciones (enlaces) y sus posiciones.

11 Problemas Y Soluciones

- Un primer problema se presento cuando intentamos almacenar el texto contenido en los controles de texto enriquecido (RitchTextBox) a la base de datos, el problema se presentaba con algunos caracteres del Formato RTF. De las soluciones que planteamos como decodificar la información en un Archivo plano, realizar reemplazos, etc. Se eligió utilizar Subsonic, un componente que actúa sobre Visual Studio .Net que simplifica y reduce trabajo de desarrollo de la capa de datos. Que para nuestro problema fue de gran utilidad, con la generación de

las entidades del Negocio que soportan el ingreso de información a la base de datos, evita la inyección de SQL, entre otros.

- Otro inconveniente se presento al querer dar persistencia a las plantillas de los procesos, debido a que estas se basan en un Proceso Base y como el objetivo final de la plantilla es podar el Proceso Base, debíamos crear un modelo que lo permitiera, sin afectar el proceso Base. Se plantearon diferentes diseños hasta llegar al actual diseño de Base de Datos, donde la Plantilla se gestiona como copia completa que esta relacionada directamente con el Proceso Base de tal forma que el Proceso Base no es afectado.

- Hacer que el WORK FLOW permitiera la adición de algoritmos no fue una tarea fácil. Esto requirió, primero, revisar la forma cómo .Net carga los Assemblies y cómo hace para presentar la información de los métodos en el Objec Inspeccion. Llevando como consecuente con la utilización del NameSpace "System.Reflection" del Framework de .Net y al diseño de interfaces que los algoritmos deben cumplir para su correcta inclusión en CMIN.

- Un problema aún no resuelto de la mejor forma, es el registro de Información en las actividades de los Proyectos a través de Plantillas RTF. El problema parte del hecho que se necesito utilizar el control de texto enriquecido (RitchTextBox) para que el usuario registre la información, pero este tiene un inconveniente con el manejo de tablas ya que no respeta los límites de las tablas y no permite adicionar más Columnas o Filas. Actualmente lo que se hace es exportar la información a un archivo con Formato RTF donde el usuario actualiza la información con una Herramienta como Word, luego este archivo es subido a CMIN nuevamente para registrar los cambios en la Base de Datos. La mejor solución a este problema es integrar un editor de Texto a CMIN que permita el correcto manejo de Tablas. Esta solución se desarrollo pero con controles en versiones TRIAL los cuales no tenían gran estabilidad así que para esta versión de CMIN se dejo con el RitchTextBox que viene incluido en .NET sin costo alguno.

12 Evaluación de la TOOL CASE CMIN

La herramienta CASE CMIN ha sido sometida a diferentes evaluaciones durante su desarrollo. Iniciando con pruebas realizadas por estudiantes de la electiva Minería de datos (del segundo periodo de 2007) realizadas en Febrero de 2008, una nueva evaluación del proyecto fue realizada en la convocatoria de proyectos para ser presentados en el Demofest Microsoft Research Academic Summit al cual se envío un Póster científico del proyecto con el cual se logro el aval para su presentación en Mayo de 2008 y finalmente se realizo una prueba Beta con ingenieros y estudiantes de Ingeniería de sistemas que trabajan en el campo de minería de datos, la cual consistió en un taller Realizado en CMIN, un experimento sencillo acerca del aprendizaje con la Herramienta y un test de usabilidad de la misma. A continuación se presentan algunos detalles de estas tres pruebas.

12.1 Evaluación de Prototipo CMIN por estudiantes de la Electiva Minería de Datos

Para esta evaluación, el prototipo de CMIN contaba con los Módulos de Gestión de Procesos y Gestión de Proyectos en una versión inicial. La evaluación consistió en revisar las Fases de la metodología CRISP en un Proyecto y verificar que estuviera todo lo propuesto en la guía de usuario de CRISP-DM.

Las fases se repartieron de la siguiente manera entre los estudiantes:

- Comprensión del Negocio: Alba Viviana Camayo y Adrián Fernando Martínez.
- Comprensión de los Datos: Oscar Eduardo Rendón y Alexander Ortiz Rosada.
- Procesamiento de los Datos: German Velasco y Jose Luis López
- Modelado: Jennifer Andrade y Willian Constain
- Evaluación: Diego Benavides y Andres Benavides.
- Despliegue**:** Deiro Enrique zuñiga y William Ramiro Joaqui

Como resultado de la evaluación los estudiantes entregaron un documento con la evaluación (Anexo D) y sugerencias.

Como conclusión de esta evaluación se verifico que CMIN contiene todas las actividades propuestas en cada Fase, pero en algunas tareas, la Plantilla de recolección de Información propuesta no era la adecuada. Además se destaco un problema en común en todas las evaluaciones con el Ingreso de Información en las Plantillas, el cual no respetaba el fin de las tablas y no se permitía adicionar más registros a las Tablas.

Estos problemas registrados se solucionaron permitiendo la Exportación de las Plantillas presentadas para el ingreso de información de la actividad a un archivo con formato RTF el cual puede ser editado en una aplicación como Word y cargado nuevamente a CMIN. De esta Forma el usuario puede mejorar las plantillas propuestas o plantear unas propias.

12.2 Presentación en el DEMO FEST Microsoft Research Academic Summit

El proyecto CMIN fue seleccionado para ser presentado en el DemoFest Microsoft Research Academic Summit realizado en Panamá del 14 al 16 de mayo de 2008. Esta presentación consistió en la exposición de un Póster del Proyecto (Anexo C) y una demostración en vivo al público del evento, de todas las funcionalidades de la TOOL CASE CMIN disponibles hasta ese momento. Aunque el evento reunió Proyectos de America Latina con Inversiones muy superiores al nuestro, CMIN logró destacarse recibiendo excelentes comentarios de los investigadores asistentes al evento y ser parte de los cinco proyectos seleccionados por Microsoft para salir en una noticia del canal CNN en el programa Adelantos (disponible en http://www.unicauca.edu.co/~ccobos/cnn-adelantos.wmv). Además recibió propuestas para que el proyecto fuere compartido en SorceForge.net y tener una comunidad de desarrollo para el mismo. Este último tema, aún esta en discusión por parte del grupo.

12.3 Prueba Beta con estudiantes que trabajan en minería de datos

Esta actividad tenía como primer objetivo, la revisión completa de CMIN en su versión final, en un ambiente real diferente al ambiente de desarrollo, lo que requirió realizar un proceso de Instalación previo a la prueba, en el cual se resolvió un inconveniente no detectado con las conexiones al servidor de Base de datos.

Otro de los objetivos de esta prueba, consistía en verificar con un experimento corto, si mediante la realización de ciertas actividades y seguimiento de un Proyecto especifico en la Herramienta CMIN se puede apoyar el aprendizaje de la Metodología CRISP-DM teniendo en cuenta que los Proyectos se Basan en dicha metodología.

Finalmente esta Prueba buscaba realizar una prueba de usabilidad, partiendo de la propuesta de un grupo de investigación de la Universitat Politecnica de Catalunya [43], para valorar el estado y aceptación del diseño de las interfaces de usuario de la aplicación. Teniendo en cuenta que este diseño fue realizado en base al conocimiento adquirido a través de la carrera, la cual en su currículo no tiene este campo como un tema central.

Para esta Prueba se reunió un grupo de usuarios que tuviesen algún conocimiento con la temática de Minería de Datos como probadores de CMIN para obtener no solo corrección de errores de codificación sino también apreciaciones sobre la funcionalidad. Se creo un test (Anexo E) que contiene preguntas acerca de la Metodología CRISP-DM sin mencionar que se trata de dicha metodología, permitiendo valorar que tanto el grupo de prueba conoce esta metodología. Parte de este test se enfoco en conocer que herramientas de minería de datos son las más conocidas y utilizadas por el grupo. Además se preparo un taller de minería de datos para la utilización del WORK FLOW de CMIN, el cual se plantea un problema de Clasificación. Dicho Problema consistió en clasificar plantas Iris entre sus diferentes Clases de Iris, Iris-cetosa, Iris-Vesicular e Iris-Virginica. Basados en información como el ancho y largo del sépalo y pétalo, de ejemplares ya clasificados. El objetivo de este taller fue clasificar en alguna de estas clases de Iris nuevos ejemplares de los cuales solo se tiene la información del ancho y largo del sépalo y pétalo.

Esta prueba se realizo de la siguiente manera:

- Aplicación del test para valorar conocimiento en CRISP-DM (ver Figura 52).
- Presentación básica de la Herramienta CMIN.
- Creación de una Plantilla del Proceso CRISP-DM pensando en el Problema del IRIS.
- Desarrollo del Taller de minería de datos (ver Figura 53).
- Presentación de los diferentes Módulos y funcionalidades de CMIN. (ver Figura 54).
- Preguntas y sugerencias (ver Figura 55).
- De Nuevo aplicación del test para valorar conocimiento en CRISP-DM.
- Aplicación del test de usabilidad.

A continuación se presentan algunas Fotos tomadas mientras se realizó la prueba Beta (ver Figuras 52, 53, 54 y 55).

Figura 52. Aplicación de Test.

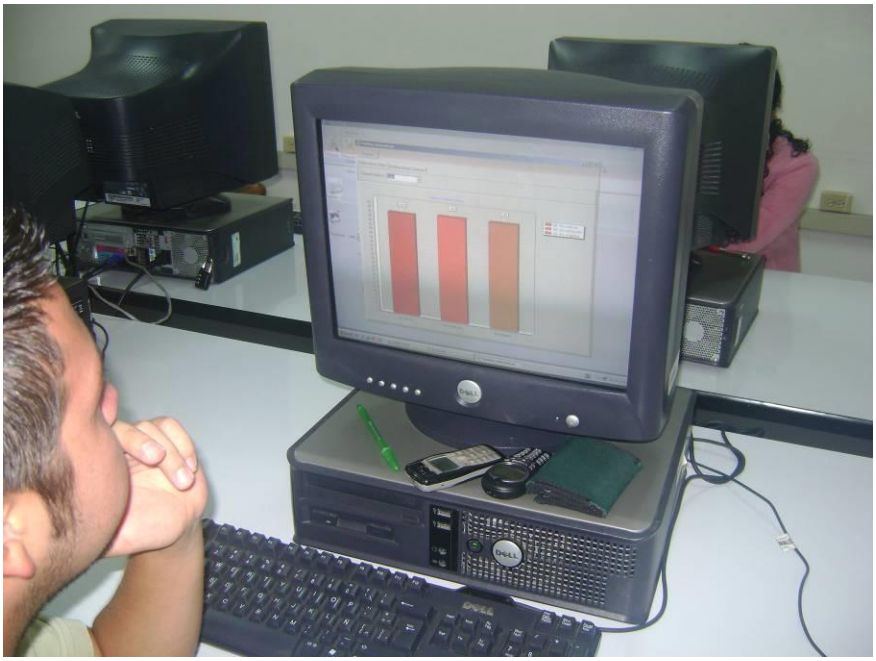

Figura 53 Desarrollo del Taller de minería y pruebas A DOC.

Figura 54. Explicación de los Módulos de CMIN.

Figura 55. Interacción con el grupo de prueba.

El grupo de prueba fue de 10 personas presentadas en la Tabla 9.

Nombre	Proyecto en que Trabaja
Diana Sánchez	Reconocimiento Balístico
Jennifer Andrade	Clustering Documentos Web
Willian Constain	Clustering Documentos Web

Oscar Rendón	Herramienta CASE para Bodegas de Datos
Germain Bolaños	Buscador Inteligente basado en minería de datos
Elvis Pérez	Buscador Inteligente basado en minería de datos
Daniel Pino	Clustering Documentos Web
Pablo Zuñiga	Clustering Documentos Web
Ing. Edwin Caldon	Minería Web en programa de Maestría
Ing. Carlos Téllez	Minería de datos genómicos

Tabla 9. Listado del Grupo de Prueba.

Resultados: Inicialmente se presenta la comparación de los resultados del Test que se aplico dos (2) veces al grupo de prueba, antes de usar CMIN y después de realizar las actividades de la prueba con CMIN. Seguido de los resultados del test de usabilidad de la Herramienta y finalmente se presenta el listado de sugerencias y problemas encontrados en esta prueba.

12.3.1 Taller IRIS de minería de Datos en CMIN

Para esta prueba los usuarios probaron el WORK FLOW de minería de Datos de CMIN, realizando un modelo de cierto grado de dificultad. Estas pruebas fueron exitosas en cuanto a no presentar errores y dada la situación que todos los usuarios resolvieron el Problema de clasificación.

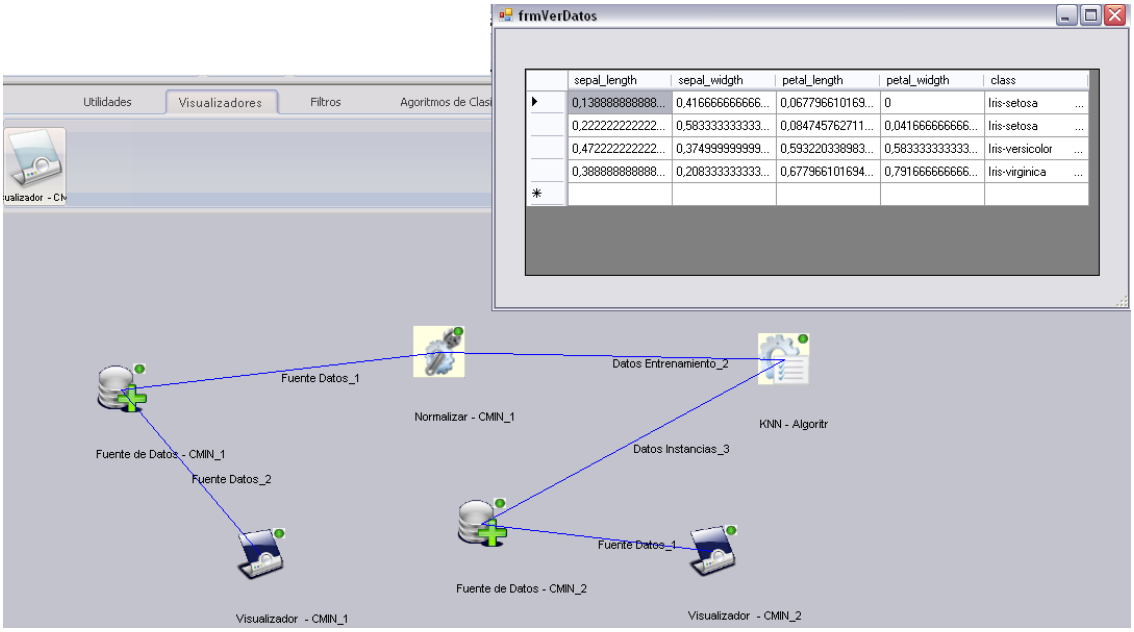

Figura 56. Modelo desarrollado en Taller Iris de minería de Datos.

12.3.2 Comparación de Test

Test 1 Antes de CMIN
Test 2 Después de CMIN

Resultados pregunta 1:
Conoce usted metodologías o procesos para el desarrollo de proyectos de minería de datos.

SI _____ NO_____

Si respondió SI escriba cuales _____

En el Test 1 un 90% de los encuestados, respondió que si conocían una metodología y a la pregunta ¿cual?, mencionaron a CRISP-DM.
En el Test 2 un 100% de los encuestados, respondió que si conocían una metodología y a la pregunta ¿cual?, mencionaron a CRISP-DM.

Resultados Pregunta 2:
Para desarrollar un Proyecto de minería de Datos cree usted necesario una etapa de "Entendimiento del negocio".

SI _____ NO_____

Si respondió SI ¿Por qué cree usted que es necesaria?
_____

Esta respuestas se evaluaron en base a la descripción de la fase de "Entendimiento del negocio" que dice "Busca comprender los objetivos del proyecto, los requerimientos del negocio, y luego convertir esta información en una definición de un problema. Incluye la realización de un plan preliminar".

En el Test 1 un 100% de los encuestados, respondió concordemente con la descripción de la fase.
En el Test 2 un 100% de los encuestados, respondió concordemente con la descripción de la fase, pero es de notar que las respuestas del Test 2 fueron mas especificas e incluyeron términos de las tareas genéricas de la fase, lo cual no se dio en el Test 1.

Resultados pregunta 3:
¿Que tareas realizaría en una etapa "Entendimiento del negocio" para un proyecto de minería de Datos?
_____

Para evaluar esta pregunta se utilizo la información presentada en la Tabla 2 donde representan las Tareas genéricas y especificas de cada Fase.

En el Test 1 un 100% de los encuestados, escribió una lista de tareas cuya definición fu acorde con las tareas planteadas en CRISP-DM para dicha fase.
En el Test 2 un 100% de los encuestados, escribió una lista de tareas cuya definición fue acorde con las tareas planteadas en CRISP-DM para dicha fase. Pero de nuevo se presento el fenómeno de que las respuestas en el Test2 incluyeron términos de las tareas genéricas de la fase, lo cual no se dio en el Test 1. Como por ejemplo la siguiente respuesta de uno de los participantes del Test. "Conocer el entorno del Proyecto: Organización, recursos con los que se cuenta, los riesgos. Conocer los objetivos del negocio, las expectativas de los clientes, conocer objetivos de minería de datos en el proyecto. Estimar el tiempo de duración del Proyecto"

Resultados pregunta 4:
¿Suponiendo las siguientes dos (2) etapas para un proyecto de minería de datos, relaciona cada tarea presentada con la etapa en la que se deba realizar?

ETAPAS		TAREAS	
A	**Entendimiento de Los Datos**		Reporte de Limpieza de Datos
			Reporte de Calidad de los Datos
			Explorar Datos
B	**Preparación de Los Datos**		Construcción de Datos

Para evaluar esta pregunta se dio un puntaje de 0 a 10, dado que hay 4 opciones, cada opción correcta tiene un peso de 2,5. Las respuestas correctas de esta pregunta fueron tomadas de la Tabla 2.

El promedio del resultado de esta pregunta en el Test 1 fue 6,75
El promedio del resultado de esta pregunta en el Test 2 fue 7,25, lo que implica una mejora en los conceptos a pesar del corto tiempo de uso de la herramienta.

Resultados pregunta 5:
Suponiendo una etapa de Modelado para Realizar un Proyecto de minería de Datos, Marque las tareas que realizaría en esta etapa.

	TAREAS
	Seleccionar la Técnica de Modelado
	Evaluar los resultados
	Evaluar el modelo
	Selección de Datos
	Generar un diseño de Test del modelo

Para evaluar esta pregunta se dio un puntaje de 0 a 10, dado que hay 3 opciones correctas, cada opción correcta tiene un peso de 3,3, pero se condiciona a que si se marca una opción errada anula una buena. Las respuestas correctas de esta pregunta fueron tomadas de la Tabla 2.

El promedio del resultado de esta pregunta en el Test 1 fue 5,96
El promedio del resultado de esta pregunta en el Test 1 fue 6,29, lo que implica una mejora en los conceptos a pesar del corto tiempo de uso de la herramienta.

Resultados pregunta 6:
De las Siguientes Etapas a presentar cuales piensa usted que deberían ser las últimas 2 Etapas de un proyecto de minería de Datos.

	ETAPAS
	Entendimiento del Negocio
	Evaluación
	Modelado
	Despliegue
	Preparación de los Datos

Para evaluar esta pregunta se dio un puntaje de 0 a 10, dado que hay 2 opciones correctas, cada opción correcta tiene un peso de 5, pero se condiciona a que si se

marca una opción errada anula una buena. Las respuestas correctas de esta pregunta fueron tomadas de la Tabla 2.

El promedio del resultado de esta pregunta en el Test 1 fue 9,5
El promedio del resultado de esta pregunta en el Test 1 fue 9,5

Conclusión general: De la pregunta 1 a la 6 se evalúo el conocimiento que el grupo de prueba tenía sobre la estructura de la metodología CRISP-DM antes de la utilización de CMIN y después. Dando como resultado, según lo visto en cada pregunta, una mejora en dichos conocimientos. El objetivo de este Proyecto no es que las personas memoricen el Proceso CRISP-DM si no que conozcan que existe una serie de pasos guía para desarrollar proyectos de minería de datos y que con la práctica y realización de proyectos en esta herramienta, apropien el Modelo identificando la necesidad de cada Fase, y estos resultados son promisorios a pesar de que son preliminares.

12.3.3 Resultados test de usabilidad

A continuación se presentan las preguntas del test de usabilidad (Anexo E) con sus resultados y una descripción de los mismos.

1) Estructura de la aplicación

a) Organización estructural: distribución de elementos estructurales de la aplicación (eje. Barras de desplazamiento, zonas de contenido, botones, etc.) es buena.

En desacuerdo __ __ __ __ __ De acuerdo

Según la opinión de los encuestados la organización estructural de CMIN es buena (ver Figura 57. Resultado Organización Estructural).

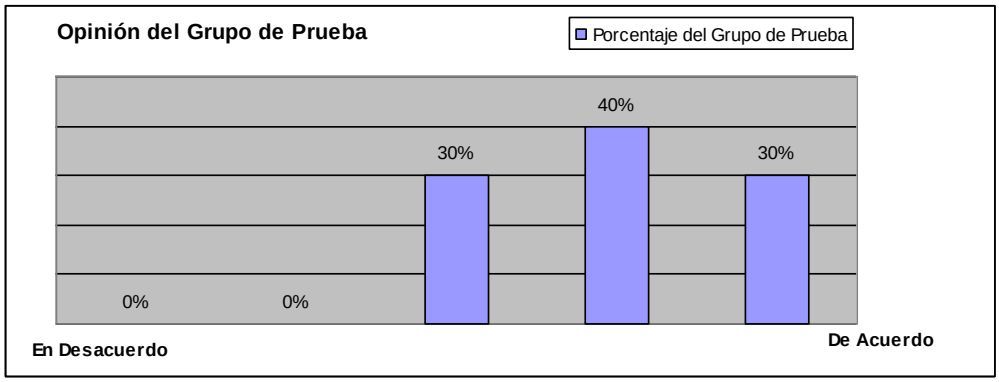

Figura 57. Resultado Organización Estructural.

b) Densidad estructural: la cantidad de elementos estructurales que se utilizan en la aplicación es excesiva.

En desacuerdo __ __ __ __ __ De acuerdo

Sobre la densidad estructural hay opiniones divididas esto debido a la cantidad de interfaces que manejaron. (Ver Figura 58. Resultados densidad estructural)

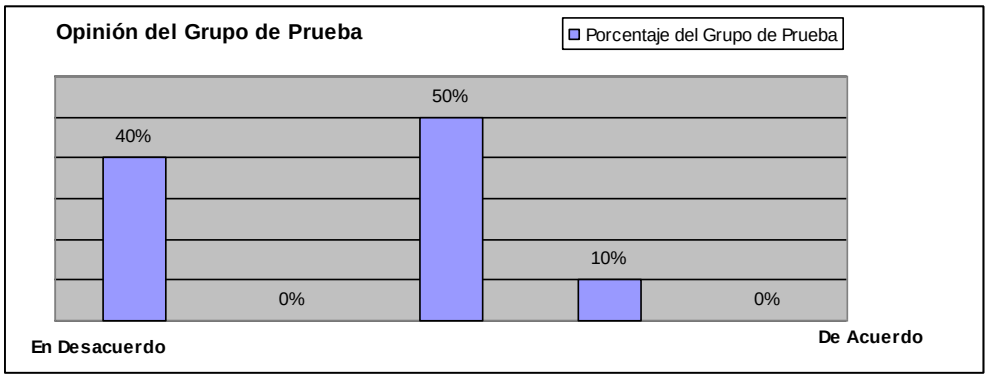

Figura 58. Resultados densidad estructural.

c) Consistencia de la estructura: la distribución de los elementos estructurales se mantiene constante a lo largo de la aplicación.
En desacuerdo __ __ __ __ __ De acuerdo

Sobre la Consistencia estructural el 100% de las opiniones están enmarcadas en los dos últimos niveles mas cercanos a la opción "De Acuerdo", Indicando buenas opiniones y era de esperarse debido a que para las ediciones de información en CMIN se mantuvo el mismo esquema. (Ver Figura 59. Resultados Consistencia estructural)

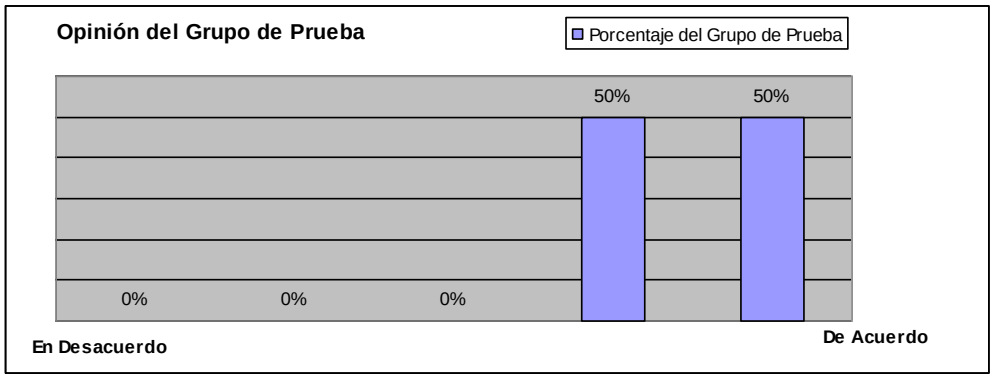

Figura 59. Resultados Consistencia estructural.

2) Operación de la aplicación

a) Navegabilidad: el recorrido que se hace por el contenido de la aplicación es fácil.
En desacuerdo __ __ __ __ __ De acuerdo

Respecto a la navegabilidad hay opiniones divididas pero la mayoría opina que es fácil hacer el recorrido sobre CMIN. (Ver Figura 60. Resultados Navegabilidad.)

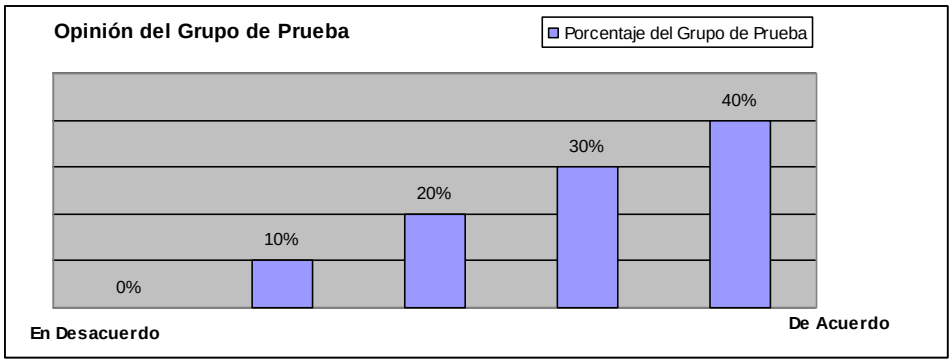

Figura 60. Resultados Navegabilidad.

b) Interactividad: la relación mutua entre el usuario y la aplicación es buena.
En desacuerdo __ __ __ __ __ De acuerdo

En cuanto a la Interactividad hay diferentes opiniones pero tienden a especificar que CMIN tiene buena interactividad con el usuario. (Ver Figura 61. Resultados Interactividad)

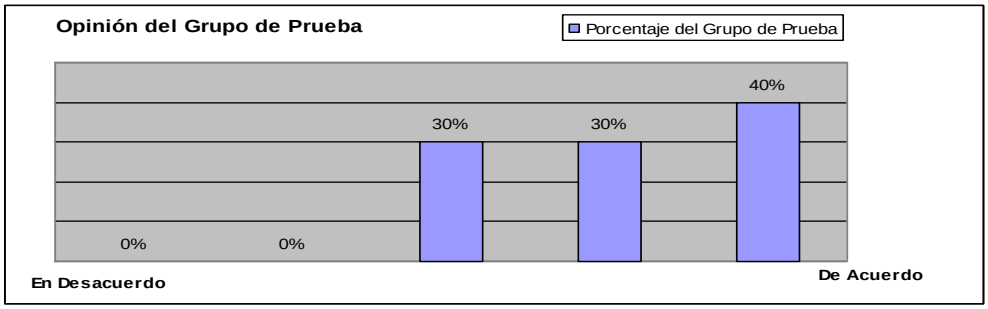

Figura 61. Resultados Interactividad.

c) Accesibilidad: las acciones que solicita la aplicación son fáciles de ejecutar.
En desacuerdo __ __ __ __ __ De acuerdo

Sumando los dos últimos niveles cercanos a la opinión "De Acuerdo" nos da que un 90% del grupo opina que las acciones que solicita CMIN son fáciles de ejecutar. (Ver Figura 62. Resultados Accesibilidad)

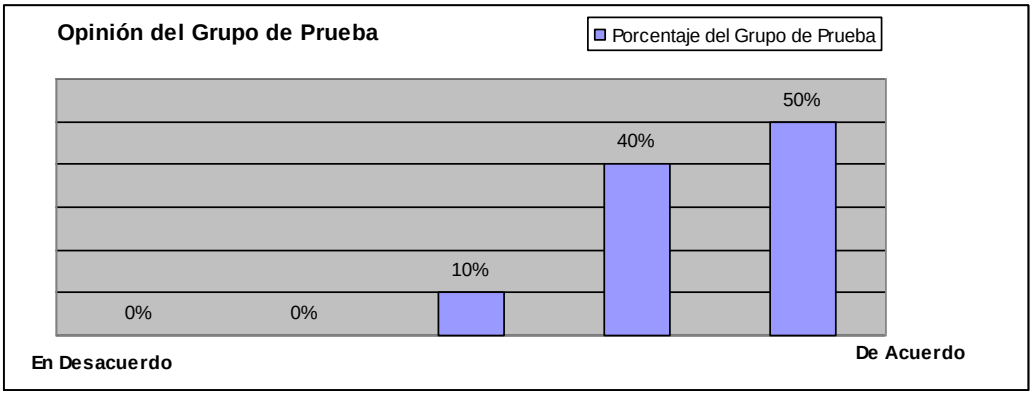

Figura 62. Resultados Accesibilidad.

d) Sistema de indicación: se identifican fácilmente las figuras, las tablas, las zonas activas y el tipo de acción que se debe ejecutar.
En desacuerdo __ __ __ __ __ De acuerdo

Teniendo en cuenta los dos últimos niveles cercanos a la opción "De Acuerdo", la gran mayoría aunque no esta totalmente de acuerdo opina que el sistema de indicación de CMIN facilita las acciones a realizar. (Ver Figura 63. Resultados sistema de indicación)

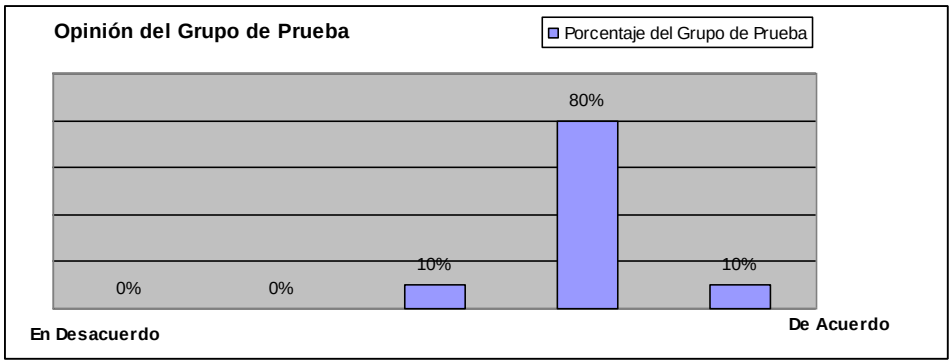

Figura 63. Resultados sistema de indicación.

e) Desempeño del sistema: la velocidad de funcionamiento de la aplicación, considerando el tipo e tarea que se exige, es buena.
En desacuerdo __ __ __ __ __ De acuerdo

Sumando los dos últimos niveles cercanos a la opinión "De Acuerdo" nos da que un 100% del grupo opina que el desempeño de CMIN es bueno.
(Ver Figura 64. Resultados Desempeño del sistema)

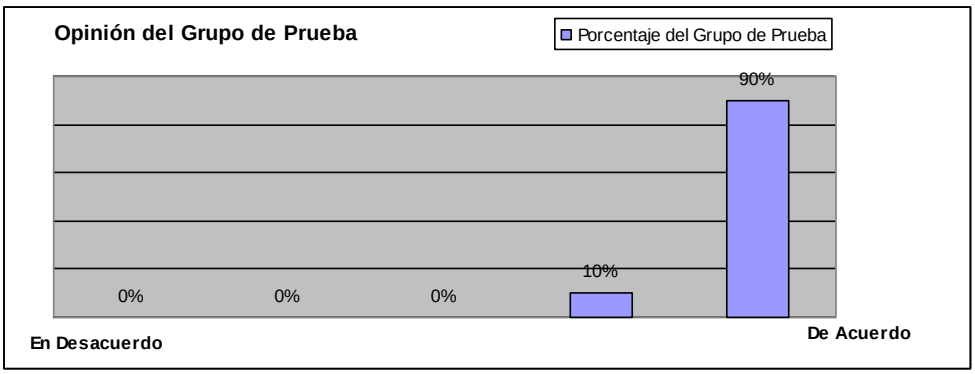

Figura 64. Resultados Desempeño del sistema.

f) Fiabilidad del sistema: hay demasiados errores durante la operación de la aplicación.
En desacuerdo __ __ __ __ __ De acuerdo

En cuanto a la fiabilidad del sistema hay opiniones divididas debido a que a algunos se les presentaron algunos problemas. (Ver Figura 65. Fiabilidad del Sistema)

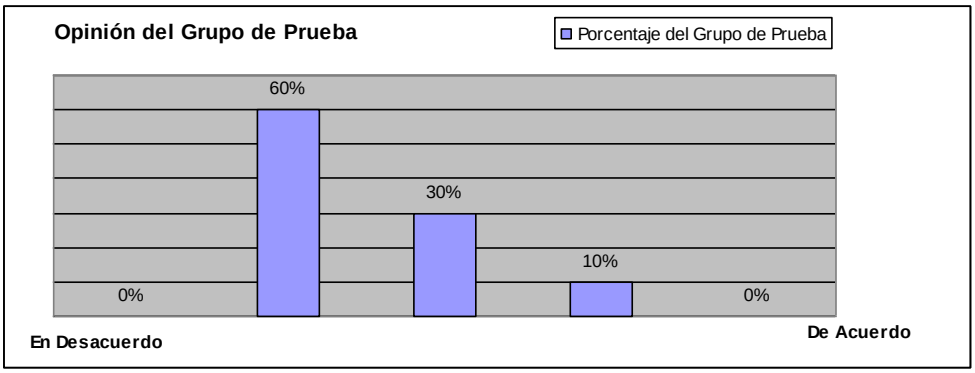

Figura 65. Fiabilidad del Sistema.

g) Consistencia de la aplicación: la ejecución de tareas (eje. Navegar por la aplicación, hacer clic en botones, seleccionar opciones etc.) sigue un estándar a lo largo de la aplicación.
En desacuerdo __ __ __ __ __ De acuerdo

Teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo" la gran mayoría opina que la consistencia en CMIN es buena, realmente ese fue uno de nuestros principios para el desarrollo de las interfaces de usuario de CMIN. (Ver Figura 66. Resultados Consistencia de la)

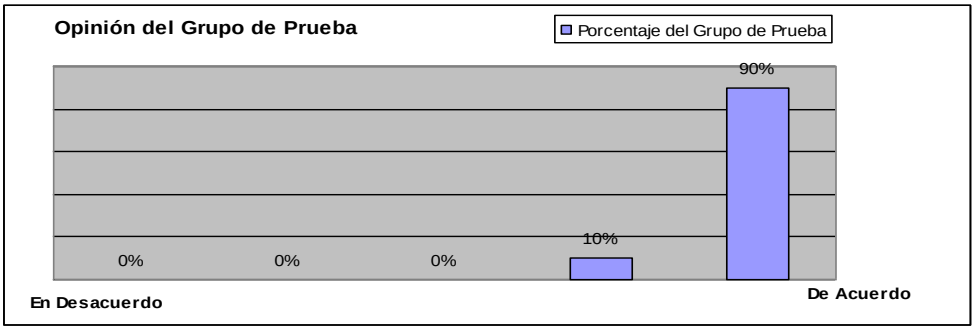

Figura 66. Resultados Consistencia de la Aplicación.

3) Información al usuario.
a) Sistema de ayuda: las dudas del usuario se resuelven fácilmente.
En desacuerdo __ __ __ __ __ De acuerdo

Hay opiniones dividas pero teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo" un mayor porcentaje opina que el sistema de ayuda es bueno en CMIN. (Ver Figura 65. Fiabilidad del Sistema.)

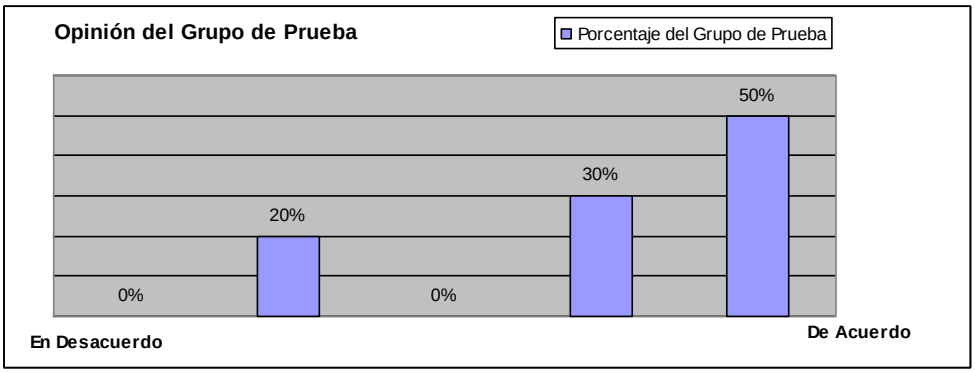

Figura 67. Resultados Sistema de ayuda.

b) Feedback (realimentación): la aplicación mantiene al usuario informado sobre las tareas en ejecución.
En desacuerdo __ __ __ __ __ De acuerdo

Hay opiniones dividas pero teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo", la mayoría del Grupo de prueba opina que CMIN los mantiene informados cuando se realizan Tareas. (Ver Figura 68. Resultados Realimentación.)

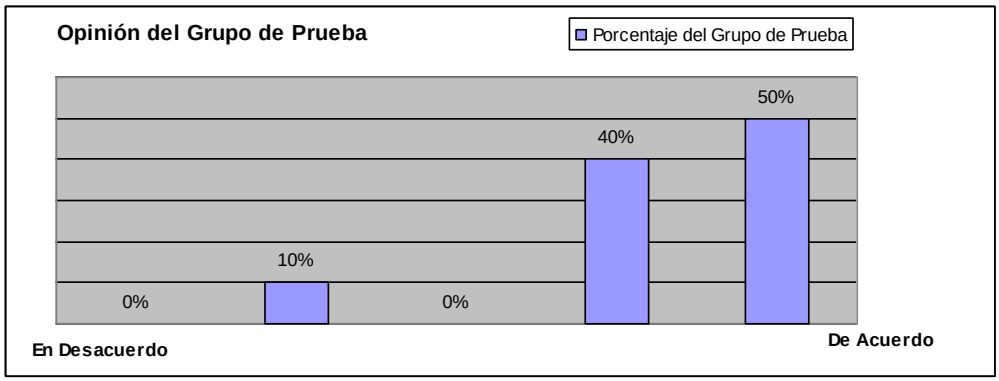

Figura 68. Resultados Realimentación.

c) Búsqueda de información: los datos que busca el usuario son fáciles de encontrar.
En desacuerdo __ __ __ __ __ De acuerdo

Hay opiniones divididas pero se debe notar que para ser la primera vez que se utilizo CMIN los usuarios no tuvieron muchas dudas en los pasos a realizar. (Ver Figura 69. Resultados Búsqueda de Información)

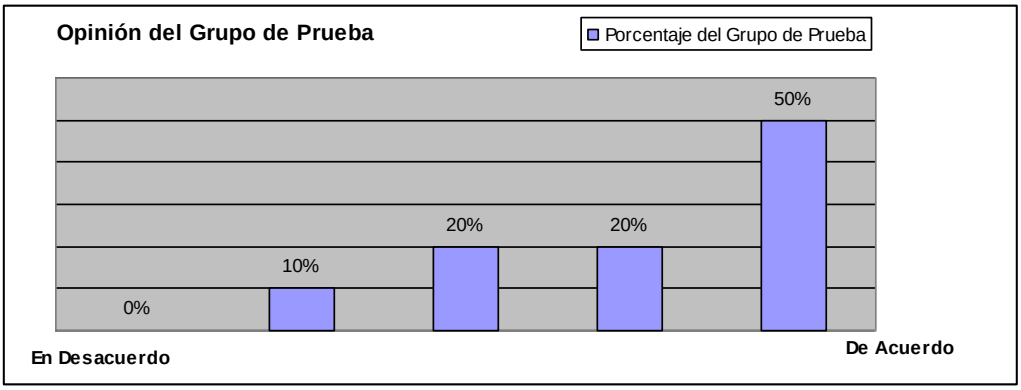

Figura 69. Resultados Búsqueda de Información.

4) Apariencia: la presentación del contenido (eje. El tipo y tamaño de fuente, e uso de color, disposición de los elementos según su significado etc.)es buena.
En desacuerdo __ __ __ __ __ De acuerdo

Hay opiniones divididas pero teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo", la gran mayoría opina que la apariencia de CMIN es buena. (Ver Figura 70. Resultados Apariencia)

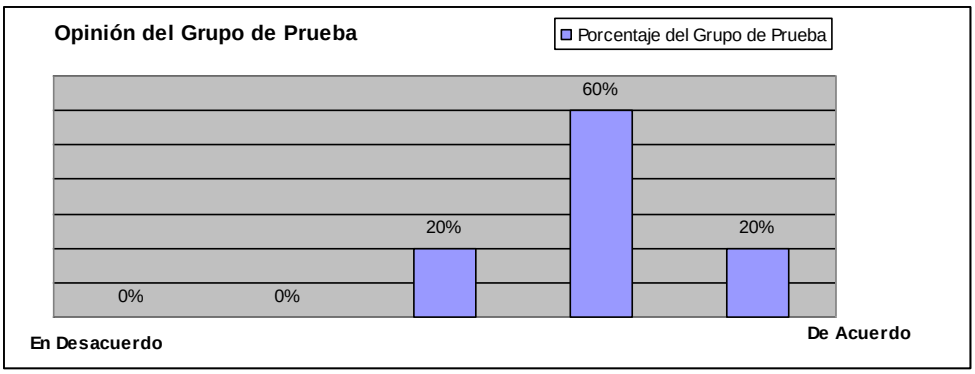

Figura 70. Resultados Apariencia.

5) Intuición: los procedimientos de navegaron por la aplicaron o ejecución de las tareas asignadas se aprenden de forma prácticamente inmediata.
En desacuerdo __ __ __ __ __ De acuerdo

Según la opinión de los encuestados CMIN es intuitivo y las tareas a realizar en ella son fáciles de aprender. (Ver Figura 71. Resultados Intuición)

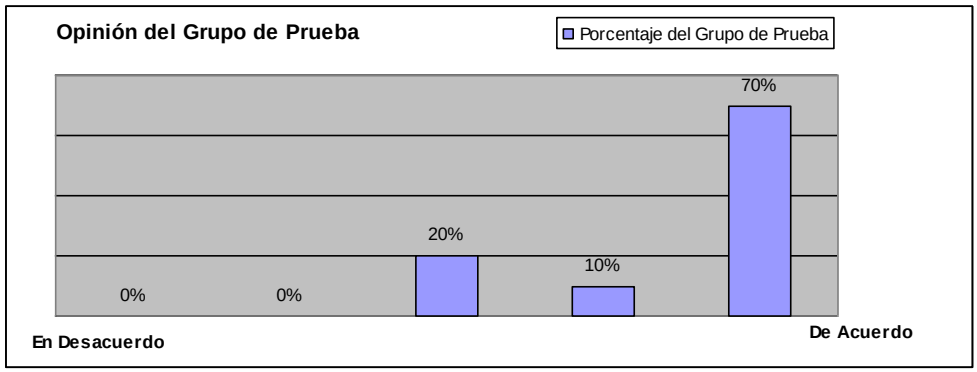

Figura 71. Resultados Intuición.

6) Contenido

a) Organización de contenido: la distribución del contenido de la aplicación (eje. Textos, imágenes, test, etc.) es buena.
En desacuerdo __ __ __ __ __ De acuerdo

Teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo", un gran porcentaje del grupo de prueba opina que la Organización del Contenido en CMIN es buena. (Ver Figura 72. Resultados Organización de Contenido)

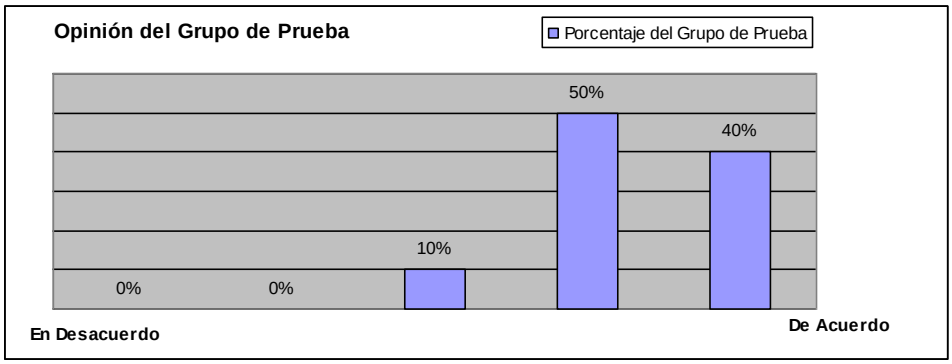

Figura 72. Resultados Organización de Contenido.

b) Densidad de contenido: la información que se presentan en la aplicación es demasiado extensa.
En desacuerdo __ __ __ __ __ De acuerdo

Teniendo en cuenta los 2 niveles mas cercanos a "De Acuerdo", la gran mayoría del grupo de prueba opina que la información se presenta de manera adecuada. (Ver Figura 73. Densidad del Contenido)

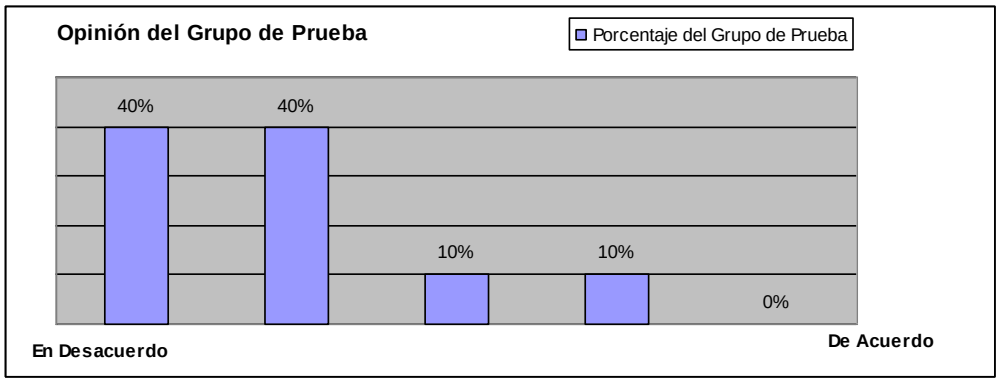

Figura 73. Densidad del Contenido.

c) Fiabilidad del contenido: no hay errores en la información que se presenta en la aplicación.
En desacuerdo __ __ __ __ __ De acuerdo

En cuanto a los resultados que arroja CMIN hay opiniones divididas en el grupo de Prueba, pero más de la mitad de los encuestados opina que no hay errores en la información que presenta CMIN. (Ver Figura 74. Resultados Fiabilidad del Contenido)

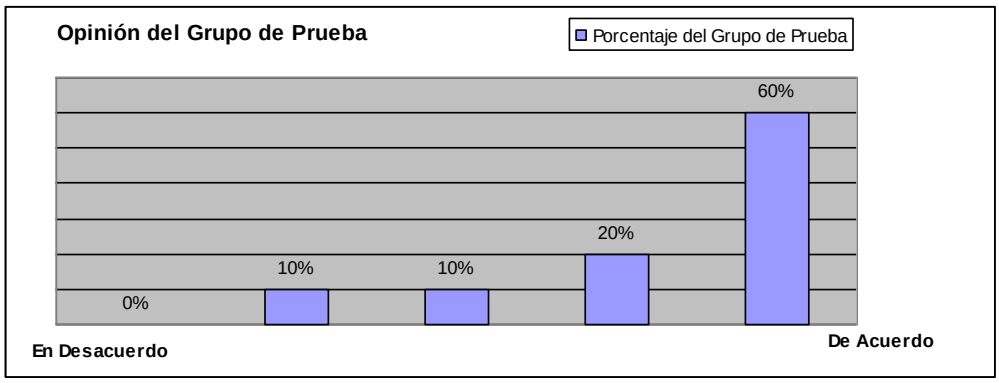

Figura 74. Resultados Fiabilidad del Contenido.

d) Comprensión del contenido: la información que se presenta en la aplicación es fácil de entender y memorizar.
En desacuerdo __ __ __ __ __ De acuerdo

Teniendo en cuenta los dos últimos niveles del lado "De Acuerdo", una gran mayoría del grupo de prueba piensa que fácil la comprensión de la información contenida de CMIN. (Ver Figura 75. Resultados Comprensión del Contenido)

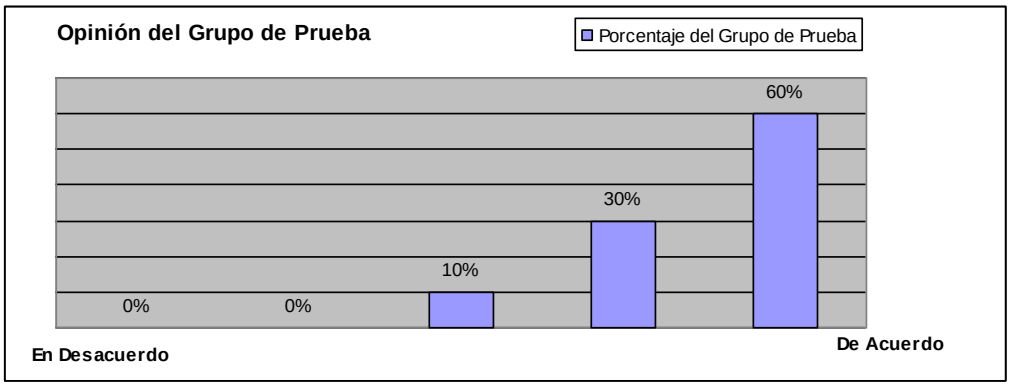

Figura 75. Resultados Comprensión del Contenido.

7) Experiencia del usuario **Prescindible __ __ __ __ __ Imprescindible**

Hay opiniones divididas en cuanto a si la experiencia del usuario es necesaria para la utilización de CMIN. Esto se debe a que la gran mayoría del grupo de usuario no había utilizado herramientas con WORK FLOW y menos aun de minería de datos. Pero dada esta condición del grupo se obtuvieron CMIN obtuvo buenos resultados. (Ver Figura 76. Resultados Experiencia de Usuario.)

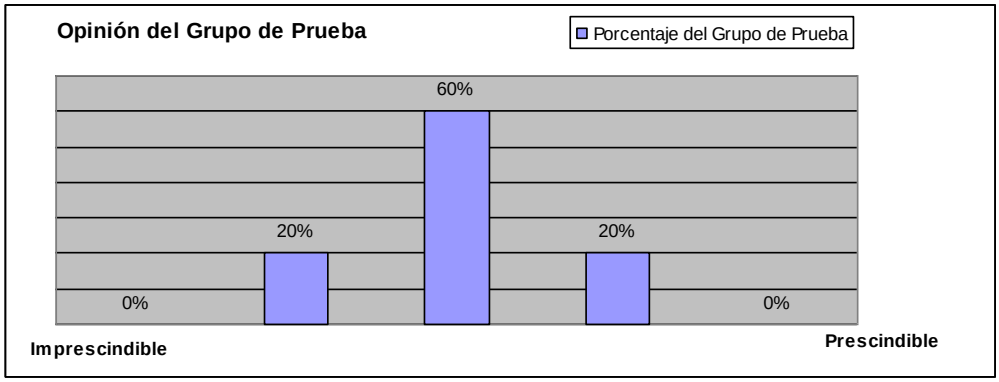

Figura 76. Resultados Experiencia de Usuario.

Opinión general sobre la aplicación. **Mala** **__ __ __ __ __ Excelente**

En términos generales se obtuvieron buenos resultados, el trabajo realizado por el grupo de investigación en este aspecto es un gran logro, ya que se puede afirmar que CMIN cuenta con una interfaz amigable, entendible y sobre todo que el manejo de los proyectos, que contemplan aspectos repetitivos y en cierta medida complejos pueden ser manejados con facilidad. La interfaz minimiza lo que el usuario debe aprender y en cada paso lo orienta para llevar a feliz término cada una de las tareas de un proyecto de minería de datos. (Ver Figura 77. Resultados opinión General sobre la Herramienta)

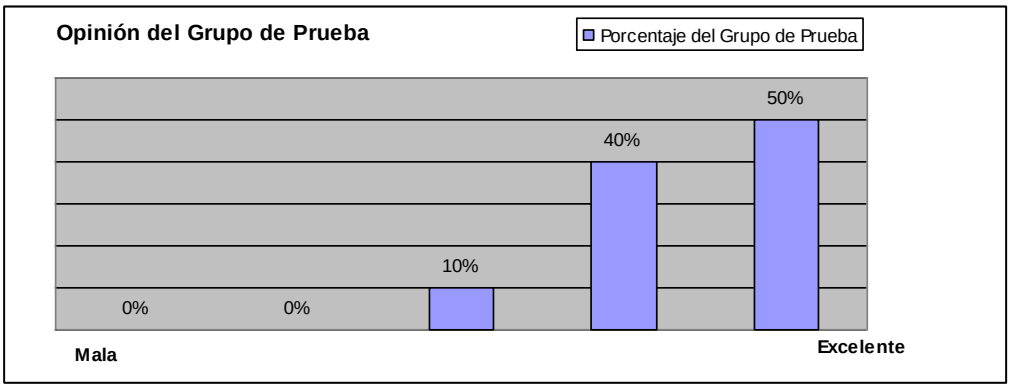

Figura 77. Resultados opinión General sobre la Herramienta.

12.3.4Sugerencias del grupo de prueba y Problemas encontrados

Sugerencias

- En la creación de plantillas sería bueno contar con la opción de deshacer cambios (restaurar pasos eliminados).
- Incluir una ayuda explicativa de los tipos de paso.
- En la descripción de los pasos se debería organizar el texto en párrafos para una lectura más agradable.
- Colocar un Icono especial y diferenciador para las sugerencias que presenta CRISP-DM para algunas actividades de los proyectos.
- Usar una barra progreso al generar los reportes.
- Manejar plantillas RTF para el registro de la información en todas las actividades.
- Cambiar el Icono que representa el manejo de fechas de las actividades.
- Disminuir el tamaño de los BOTONES – Iconos.

Problemas encontrados

Algunos de los Problemas encontrados fueron:

- En el proceso de mapeo de Procesos a plantillas no se debe permitir modificar el tipo de paso.
- Al crear los proyectos el formulario que presenta los pasos del proyecto esta saliendo descentrado en la interfaz.
- Validar que las fechas finales que se definen para el seguimiento no sean menores a las iniciales.
- Incluir barras de Progreso para algunos Procesos
- Incluir la opción deshacer al eliminar pasos en la personalización de los Procesos (Creación de Plantillas).

Todos estos Problemas fueron resueltos para la versión definitiva.

13 Conclusiones, Recomendaciones y Trabajo Futuro

13.1 Conclusiones

- Se logro modelar e implementar una Herramienta CASE integrada que orienta el desarrollo de los proyectos a través de Procesos, facilita la integración del Proceso o Metodología propuesta con el Proyecto y asegura el cumplimiento del Proceso en la ejecución del Proyecto.

- CMIN una herramienta con funcionalidad extensible, que facilita el desarrollo en comunidad, con la razón de que cada nueva funcionalidad es programada por aparte (por los miembros de la comunidad), después es probada y evaluada por un grupo de expertos y finalmente es incluida y expandida a los demás miembros de la comunidad.

- Mediante la información, detallada y apropiada del Proceso guía, que presenta CMIN en cada Paso del proyecto y la realización de actividades en esta CASE, posibilita que el usuario apropie el Proceso guía identificando la necesidad de cada Paso.

- Incluir al inicio de la fase de Planeación [41] actividades para entender y comprender el contexto del proyecto, posibilita que el grupo de desarrollo comparta sus conocimientos y experiencias con respecto al tema y de paso asegura que todos los participantes del proyecto conocen dicho contexto, lo que facilita el desarrollo del mismo.

- El grado de aceptación de CMIN como herramienta para desarrollar Proyectos de minería de datos es alto debido a los buenos comentarios recibidos en los eventos en donde ha sido presentada y las buenas opiniones de sus primeros usuarios reales en las pruebas Beta realizadas.

- CMIN es una herramienta CASE que permite planear las fechas para el desarrollo de las actividades en los proyectos, genera reportes para revisar el cumplimiento de las mismas y permite el registro de los resultados de todas las actividades de los Proyectos, posibilitando el seguimiento de los mismos.

- CMIN una herramienta que ofrece la posibilidad de utilizar su WORK FLOW de minería de datos en las actividades del Proyecto que lo requieran, teniendo como presente que las actividades de los proyectos contienen la descripción detallada de lo que se requiere realizar en las mismas y que de forma practica cada actividad tiene su propio WORK FLOW para mantener el registro de lo realizado en la misma, hace fácil la labor de Modelado y la revisión del Proceso de Minería de datos del Proyecto.

- Realizar reuniones semanales para integrar lo desarrollado por cada uno de los integrantes del grupo, facilito el manejo de versiones del código base en las diferentes fases de construcción del proyecto.

13.2 Recomendaciones

Para el desarrollo de una Herramienta CASE se recomienda dividir su creación en diferentes Proyectos, por ejemplo un Primer proyecto que realice el Modelo de la

aplicación teniendo en cuenta los Módulos futuros a realizar y su integración, segundo, proyectos que creen Módulos autosuficientes que utilicen la base del Primer Proyecto para exponer servicios que faciliten una futura integración y finalmente un tercer Proyecto que se encargue de integrar estos Módulos desarrollando lo que sea necesario para el cumplimiento del modelo planteado y los objetivos de la CASE.

13.3 Trabajo Futuro

- Implementar un componente de seguimiento a Proyectos basado en lo ya hecho que tenga en cuenta la administración de los recursos para cada actividad, de tal forma que se puedan hacer reportes de costos en cada paso del Proyecto y tener integrado en CMIN una metodología de Gestión de Proyectos en conjunto con los Procesos de Minería de datos.

- Dado que Existen unas técnicas de minería de datos que realizan combinación de Modelos (como Boosting y Baggin) [44], estas técnicas no fueron contempladas en CMIN para esta versión, el trabajo sería crear una Interfaz para este tipo de técnicas e incluir la programación necesaria para su funcionamiento en el WORK FLOW de CMIN para una futura versión.

- Crear una comunidad que permita un rápido crecimiento de los algoritmos que ofrece CMIN en la actualidad, para potenciar el uso del WORK FLOW.

- Para la continuidad del proyecto se han planteado 2 opciones: la primera es presentar la herramienta a una empresa de desarrollo software que pueda estar interesada en continuar con su desarrollo y distribución, como segunda opción es subir el código fuente a SourceForget.net el cual es una comunidad de desarrollo apoyada por Microsoft en la cual los participantes colaboran en el mejoramiento de los códigos fuentes subidos a esta.

14 BIBLIOGRAFÍA Y REFERENCIAS

[1] Ordonez, C. 2006. Comparing association rules and decision trees for disease prediction. In *Proceedings of the international Workshop on Healthcare information and Knowledge Management* (Arlington, Virginia, USA, November 11 - 11, 2006). HIKM '06. ACM Press, New York, NY, 17-24. DOI= http://doi.acm.org/10.1145/1183568.1183573

[2] Lo, V. S. 2002. The true lift model: a novel data mining approach to response modeling in database marketing. *SIGKDD Explor. Newsl.* 4, 2 (Dec. 2002), 78-86. DOI= http://doi.acm.org/10.1145/772862.772872

[3] Abdullah, A., Brobst, S., Pervaiz, I., Umer, M., and Nisar, A. 2004. Learning dynamics of pesticide abuse through data mining. In *Proceedings of the Second Workshop on Australasian information Security, Data Mining and Web intelligence, and Software internationalisation - Volume 32* (Dunedin, New Zealand). J. Hogan, P. Montague, M. Purvis, and C. Steketee, Eds. ACM International Conference Proceeding Series, vol. 54. Australian Computer Society, Darlinghurst, Australia, 151-156.

[4] Timarán, R. Arquitecturas de integración del proceso de descubrimiento de conocimiento con sistemas de gestión de bases de datos: un estado del arte, en revista ingeniería y competitividad. Revista de Ingeniería y Competitividad, Universidad del Valle, 3(2), Diciembre 2001.

[5] Timarán, R. and Millán, M. Equipasso: an algorithm based on new relational algebraic operators for association rules discovery. In Fourth IASTED Internacional Conference on Computational Intelligence, Calgary, Alberta, Canada, July 2005. http://www.actapress.com/PaperInfo.aspx?PaperID=21110

[6] Timarán, R. and Millán, M. Equipasso: un algoritmo para el descubrimiento de reglas de asociación basado en operadores algebraicos. In 4ˆAa Conferencia Iberoamericana en Sistemas, Cibernética e Informática CICI 2005, Orlando, Florida, EE.UU., Julio 2005. http://www.iiisci.org/cisci2005/program/html/program.htm

[7] Universidad Nacional de Colombia sede Manizales, (Visitado 2007, Marzo 28). Semilleros de investigación [Pagina de Información]. URL http://www.manizales.unal.edu.co/dep_infcom/semilleros.php

[8] Universidad de Antioquia, (Visitado 2007, Marzo 28). Anexos tesis de especialización [Documento de Información]. URL http://electronica.udea.edu.co/departamento/acreditacion/Anexos_archivos/ANEXO%2021%20TESIS%20DE%20ESPECIALIZACION%20EN%20CIENCIAS%20ELECTRONICAS%20E%20INFORMATICA.pdf

[9] Orientevirtual, (Visitado 2007, Marzo 29). Universidad-empresa: crece en Latinoamérica [Noticia]. URL http://www.orientevirtual.org/?2,4979,es

[10] Contraloría General de la Nación, (Visitado 2007, Marzo 29). El Contralor General de la República, Julio César Turbay Quintero, puso en marcha 15 estrategias contra la corrupción [Boletín de Prensa]. URL http://www.contraloriagen.gov.co:8081/internet/cartelera/Archivos/2924/info_noticia.jsp?id=2924

[11] Goebel, M. and Gruenwald, L. 1999. A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl.* 1, 1 (Jun. 1999), 20-33. DOI= http://doi.acm.org/10.1145/846170.846172

[12] SPSS Andino, (Visitado 2007, 30 Marzo). SPSS Inicia operaciones en Colombia [Noticia].URL

http://www.spss.com/la/colombia/press/SPSS%20Andino%20llego%20a%20Colombia.pdf

[13] Gondar Nores, José Emilio. (Visitado 2009, Marzo 21). Metodologías para la Realización de Proyectos de Data Mining. [Articulo en Inrernet]. URL http://www.estadistico.com/arts.html?20040426

[14] Project CRISP-DM (CRISP-DM). (Visitado 2009, Marzo 21). [Pagina Principal]. URL http://www.crisp-dm.org/

[15] SAS (SEMMA). (Visitado 2009, Marzo 21). [Pagina de Información]. URL http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

[16] MetaGroup (METAspectrum Market Summary), (Visitado 2009, Marzo 21), http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdf

[17] Kdnuggets (tools data mining), (Visitado 2009, Marzo 21) http://www.kdnuggets.com/polls/2005/data_mining_tools.htm

[18] SPSS Inc. (Clementine), (Visitado 2009, 21 Marzo). [Pagina de Información]. URL http://www.spss.com/es/clementine/

[19] Insightful Corporation (Insightful Miner), (Visitado 2009, Marzo 21). [Pagina de Información]. URL http://www.insightful.com/products/iminer/default.asp

[20] University of Waikato, (Visitado 2009, Marzo 21). [Pagina de principal]. URL http://www.cs.waikato.ac.nz/ml/weka/

[21] Salford System (CART), (Visitado 2009, Marzo 21). [Pagina Información]. URL http://www.salfordsystems.com/cart.php

[22] PolyAnalyst (tool data mining),(Visitado 2009, Marzo 21), http://www.megaputer.com

[23] SAS(Enterprise miner), (Visitado 2009, Marzo 21) http://www.sas.com/technologies/analytics/datamining/miner/

[24] Britos, P., Fernández, E., Ochoa, M., Merlino, H., Diez, E., y García, R. METODOLOGÍA DE SELECCIÓN DE HERRAMIENTAS DE EXPLOTACION DE DATOS. II Workshop de Ingeniería del Software y Bases de Datos. XI Congreso Argentino de Ciencias de la Computación (2005). Pág. 113-123. http://www.itba.edu.ar/capis/webcapis/RGMITBA/comunicacionesrgm/CACIC-2005-Metodologia-de-Seleccion-de-Herramientas-de-Explotacion-de-Datos.pdf

[25] Albizuri-Romero, M. B. 2000. A retrospective view of CASE tools adoption. *SIGSOFT Softw. Eng. Notes* 25, 2 (Mar. 2000), 46-50. DOI= http://doi.acm.org/10.1145/346057.346071

[26] Universidad Cooperativa de Colombia, (Visitado 2009, Marzo 21). [Pagina de principal]. URL http://www.uccpopayan.edu.co/index2.php?i=18

[27] Colegio Mayor del Cauca, (Visitado 2009, Marzo 21). [Pagina principal]. URL http://www.colmayorcauca.edu.co/programas_academicos/programa.php?p=5

[28] Fundación Universitaria de Popayán, (Visitado 2009, Marzo 21). [Pagina de principal]. URL http://www.fup.edu.co/2006/index.php?section=52

[29] Gersten, W., Wirth, R., and Arndt, D. 2000. Predictive modeling in automotive direct marketing: tools, experiences and open issues. In Proceedings of the Sixth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Boston, Massachusetts, United States, August 20 - 23, 2000). KDD '00. ACM Press, New York, NY, 398-406. DOI= http://doi.acm.org/10.1145/347090.347174

[30] INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA (Perú), Elaborado por la Sub-Jefatura de Informática. (Visitado 2006, 28 Noviembre).

Herramientas CASE [Documento Informativo]. URL http://www.inei.gob.pe/biblioineipub/bancopub/Inf/Lib5103/Libro.pdf

[31] CEDS (Centro Mundial de Formación por Internet en Análisis, Diseño Informático y CASE), (Visitad0 2009, 27 Mayo). [Pagina de Información]. URL http://ceds.nauta.es/cursos/canalisis0204.htm

[32] Sarker, Ruhul A., Hussein, A., Newton, Charles S. *Heuristic and Knowledge Discovery* (Hershey, PA, USA: Idea Group Publishing, 2002, Capitulo I Pag. 2-3) **Heuristic and Optimization for Knowledge Discovery**

[33] Universidad de Regina (Computer Science Student Society). (Visitado 2009, 27 Mayo). [Pagina de Información]. URL http://www2.cs.uregina.ca/~hamilton/courses/831/notes/kdd/1_kdd.html

[34] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39, 11 (Nov. 1996), 27-34. DOI= http://doi.acm.org/10.1145/240455.240464.

[35] Project CRISP-DM (CRISP-DM). (Visitado 2006, 01 Noviembre). [Documento]. URL http://www.crisp-dm.org/CRISPWP-0800.pdf

[36] Aluja, Tomás. (Visitado 2009, 27 Mayo). La Minería de Datos, Entre la Estadística y la Inteligencia Artificial. [Articulo en Internet]. URL http://www.idescat.cat/sort/questiio/questiiopdf/25.3.4.Aluja.pdf

[37] Peter Caron, Colin Shearer, P. 1997 Using Interactive Visual Workflow Techniques to Optimize the CRISP-DM Data Mining Process. Business Intelligence Journal. URL http://www.tdwi.org/research/display.aspx?ID=6046

[38] Interface Statement (Visual Basic), Visual Basic Language Reference (visitado 2008, Noviembre) URL http://msdn.microsoft.com/en-us/library/h9xt0sdd.aspx

[39] Reflection Overview, .Net FrameWork Developer's Guide (visitado 2008, Noviembre) URL http://msdn.microsoft.com/en-us/library/f7ykdhsy.aspx

[40] XML Web Services Basics. Web Services Technical Articles (visitado 2009, Enero) http://msdn.microsoft.com/en-us/library/ms996507.aspx

[41] LARMAN, Craig. UML y Patrones: Introducción al Análisis y Diseño Orientado a Objetos. Prentice Hall, 1999.

[42] Mahech, Chand. 2000, Creating C# Class Library (DLL) Using Visual Studio .NET (visitado 2008, Marzo) URL http://www.c-sharpcorner.com/UploadFile/mahesh/dll12222005064058AM/dll.aspx

[43] Universitat Politecnica de Catalunya, (Visitado 2009, Abril 5) http://www.tdr.cesca.es/TESIS_UPC/AVAILABLE/TDX-0716102-102210//12ApendiceA.pdf

[44] Moujahid, Abdelmalik., Inza, I., Larrañaga, P. Combinación de Clasificadores, (Universidad del País Vasco). URL http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t15s-combinacion.pdf