BENEMÉRITA

UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

“SISTEMA PARA LA TOMA DE DECISIONES PARA INCENDIOS FORESTALES DEL ESTADO DE TLAXCALA”

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN INGENIERA EN CIENCIAS DE LA COMPUACION

PRESENTA

MARIBEL FORTIZ FLORES

ASESOR

DRA MARÍA JOSEFA SOMODEVILLA GARCIA

PUEBLA, MÉXICO **OCTUBRE 2012**

Agradecimientos

Una página me es insuficiente para agradecer a todas y cada una de las personas, que de alguna u otra manera fueron partícipes de mi preparación universitaria, o que simplemente preguntaron cómo iba la tesis, a todas las personas que han hecho posible este proyecto. Para todos un gran y sincero agradecimiento.

En primer lugar agradezco a mi asesora, Dra. María Josefa Somodevilla por brindarme su apoyo, confianza y guía durante el proceso de investigación y escritura de esta tesis.

A la Comisión Nacional Forestal del estado de Tlaxcala (CONAFOR), por la confianza depositada al proporcionarme la información necesaria para el desarrollo de esta tesis. Sin ello no hubiese sido posible este logro.

A Inés Flores, mi amiga y madre, por tus sabios consejos y apoyo incondicional, a mi padre Juan Fortiz y hermanos Misael, Alberto y Fredy por apoyarme de forma incondicional y por su constante apoyo para hacer realidad cada proyecto en mi vida, gracias por animarme a seguir adelante. Finalmente a Areli Danae mi sobrina quien es mi fuente de motivación para ser cada día mejor.

Este trabajo es para todos ustedes.

Resumen

Los incendios forestales tienen un impacto negativo sobre todos los componentes del ecosistema, tanto sobre el medio físico, como biológico y humano. Cuando el monte arde, la pérdida de calidad paisajística es la consecuencia más fácilmente apreciable por la desaparición de la cubierta vegetal. Pero los incendios forestales son más destructivos y dañinos de lo que se puede observar a simple vista, afectan negativamente a todos los integrantes del ecosistema, incluido el ser humano, y sus consecuencias superan el ámbito local del terreno quemado.

El proyecto que se presenta a continuación titulado "Sistema para la Toma de Decisiones para Incendios Forestales del Estado de Tlaxcala", aborda el problema de los incendios forestales en el estado centrándose en la obtención de información adicional de utilidad en la toma de decisiones orientada a la prevención y extinción.

El planteamiento de este trabajo responde al interés de generar información de apoyo y complemento a las actividades de prevención y extinción llevadas a cabo por CONAFOR. En este sentido, la metodología que se propone puede facilitar el análisis de la información recogida por CONAFOR. La metodología consta de dos etapas fundamentales, una la diseño y creación de un Almacén de Datos en la que partiendo de unos datos iníciales se seleccionan, limpian y transforman, para finalmente sean almacenados en un Almacén de Datos que posteriormente será utilizada para la fase de análisis, en la cual se aplican técnicas de minería de datos mediante WEKA, para extraer información nueva, novedosa y útil.

En la fase de diseño del Almacén de Datos, se debe elegir en base a la información que se desea explorar, los datos que se guardaran, el hecho central, la unidad mínima de estos, las dimensiones que se estudiarán, determinar y refinar las medidas y atributos necesarios para los hechos y las dimensiones. Enseguida se procede a cargar el Almacén de Datos, mediante el sistema ETL (extracción, transformación y cargado de datos), finalmente se crea un proceso de mantenimiento para el Almacén de Datos.

Una vez preparados los datos, se aborda la fase de análisis mediante el software de libre distribución WEKA, elaborado por la universidad de Waikato, Nueva Zelanda.

En esta tesis se presentan los resultados obtenidos al aplicar técnicas de Minería de Datos a un conjunto de registros proporcionados por la Comisión Nacional Forestal del Estado de Tlaxcala (CONAFOR), que validan la utilización de la herramienta de Minería de Datos para el manejo de grandes Bases de Datos. La técnica de agrupamiento a través del algoritmo *SimpleKMeans*, indica las principales causas que generan los incendios, los meses con mayor incidencia, así como los municipios y vegetación afectada. Otra técnica aplicada es la clasificación mediante al algoritmo J48, con el cual se puede apreciar la superficie forestal que se ve afectada para cada municipio en los distintos meses del año. Finalmente se generaron reglas de asociación las cuales indican que en los municipios con mayor incidencia de incendios, su superficie afectada es pequeña comparada con la de otros municipios. Esta información es de gran importancia para la prevención y sofocación de incendios en el estado de Tlaxcala, ya que con estos datos se pueden diseñar campañas de prevención en los municipios más afectados, así como crear planes de contingencia y estrategias para la sofocación de incendios.

Índice General

Índice de Figuras

Índice de Tablas

1

Introducción

En este capítulo se presenta el contexto en el que se ubica este proyecto de tesis, el problema que se aborda, las razones que lo motivaron, la propuesta de solución y se definen los objetivos y alcances.

1.1 Introducción

Los incendios forestales son fuentes potenciales de contaminantes atmosféricos que deben ser considerados al intentar correlacionar las emisiones con la calidad del aire. El tamaño e intensidad de un incendio forestal dependen directamente de variables como: condiciones climatológicas, tipos de vegetación, grado de humedad y carga de combustible consumido por unidad de área.

Ante las nuevas tecnologías, se puede disponer de gran cantidad de información histórica, la cual puede ser utilizada para plantear estrategias de trabajo y tomar decisiones, es decir, tomar la información como una materia prima de la cual se pueda extraer conocimiento e información útil y novedosa. Dentro de estas herramientas actuales se pueden considerar los *DataWarehouse* (Almacenes de Datos).

Un *Datawarehouse* es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La información almacenada es fiable y homogénea. Se necesita de una herramienta para la toma de decisiones basándose en información integrada y global, que facilite la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén, proporcione la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios. En concreto tener una herramienta que ofrezca una optimización tecnológica y económica en entornos de centro de información, estadística o de generación de informes.

En esta tesis se desarrolla un *DataWarehouse* para tomar decisiones con respecto a los incendios forestales, el cual posteriormente será explotado por medio de técnicas de minería de datos y operaciones OLAP (*Online Analitical Processing*, por sus siglas en inglés). La información utilizada para la realización de dicho proyecto fue proporcionada por la Comisión Nacional Forestal del estado del Tlaxcala. Esta información se encuentra almacenada en documentos de Excel. Dicha información es histórica, es decir, representa hechos que se han producido a lo largo de los años.

Este proyecto conlleva dos retos para la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, entre otros). Y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. La combinación de las herramientas antes mencionadas facilita y acelera el procesamiento de la información obteniendo información clara, concisa y confiable.

1.2 Planeamiento de la investigación

En esta sección se describe el problema a resolver, se precisan los objetivos que se desean alcanzar, así como el planteamiento de la solución propuesta para dicho problema, para finalmente describir la organización de la tesis.

1.2.1 Problema a resolver

Los datos analizar en este proyecto de investigación, son proporcionados por la Comisión Nacional Forestal del estado de Tlaxcala (CONAFOR).

CONAFOR lleva un control sobre los incendios forestales que se han generan en el estado. Dichos datos son almacenados en archivos de Excel, en los que podemos encontrar fecha incendio, coordenadas geográficas, municipio, predio, paraje, causa, horario de incendio, superficie afectada (arbolado renuevo, arbolado adulto, no arbolado matorrales y arbustos y no arbolado pastizales) y personal participante (SEMARNAP, SEDENA, CGE, otras instituciones y voluntarios).

Este proyecto tiene como base, la información que ha sido almacenada por CONAFOR, dicha información recabada presenta problemas, los cuales son mencionados a continuación:

- Los datos proceden de hojas de Excel, por la cual existe inconsistencia en los tipos de datos, ruido, datos ausentes, intratabilidad, volatilidad de los datos, entre otros.

- Actualmente los datos son analizados e interpretados manualmente, invirtiendo en ello gran cantidad de horas de trabajo. Con lo que concluimos que esta forma de trabajo es pesada, cara y altamente subjetiva.

- Debido al crecimiento de las bases de datos, manejar la información de la forma tradicional se vuelve complicado, por lo que siguen la intuición del especialista al no disponer de las herramientas necesarias.

El objetivo planteado para este proyecto es ambicioso pero factible, puesto que el objetivo un *DataWarehouse* es almacenar grandes cantidades de información por otro lado, el objetivo de la Minería de Datos es obtener conocimiento novedoso y útil de bases de datos.

1.2.2 Objetivos de la investigación

El objetivo del proyecto de investigación es el siguiente:

Desarrollar un *DataWarehouse* para la toma de decisiones sobre incendios forestales en el estado de Tlaxcala, para posteriormente explotarlo mediante técnicas de Minería de Datos y operaciones OLAP.

Los objetivos particulares son los siguientes:

1. Identificar las fechas con mayor índice de incendios, así como las causas de ellos.
2. Encontrar una relación entre la zona y la causa que origina el incendio.
3. Identificar zonas con mayor índice de incendios.
4. Facilitar el lanzamiento de campañas, para la prevención de incendios.
5. Facilitar el proceso de la toma de decisiones basado en información estadística.

1.2.3 Justificación de la investigación

Los incendios forestales implican un cambio importante en los factores ecológicos que rigen el funcionamiento de los ecosistemas y dada la importancia que han adquirido en las últimas décadas, constituyen uno de los problemas ecológicos mas graves a los que han de enfrentarse los gestores del medio ambiente en nuestro país.

Ante los problemas y las dudas a las que se enfrentan los gestores del medio ambiente, se busca la utilización de nuevas infraestructuras de comunicación con potentes y flexibles herramientas de tratamiento de información (bases de datos, *Data Warehouse* (DW), *Data Mining*.) que mejoran la calidad, cantidad y eficiencia de los datos, así como el análisis, procesamiento y comunicación de los mismos. En otras palabras, pueden aportar a la institución la base tecnológica necesaria para afrontar los nuevos retos de la situación actual y las perspectivas a futuro. De ahí, que en este trabajo, se resalte el hecho de que las bases de datos y el DW permiten en primera instancia el almacenamiento adecuado de los datos obtenidos de las actividades habituales de organización, producción, control de gestión, planificación estratégica etc.. Pero, además se incide en otro hecho, que es el que a través de dichas herramientas la institución puede extraer de dichos

datos, la información y el conocimiento que necesitan para identificar y responder estratégicamente sus necesidades y retos que enfrentan.

1.3 Presentación de la solución

La propuesta para resolver el problema antes mencionando y las herramientas desarrolladas son explicadas en la siguiente sección.

1.3.1 Propuesta de solución

Una vez planteado el problema, se debe definir una solución, la cual se enuncia a continuación:

Se propone el diseño e implantación de un *Data Warehouse* multidimensional para la toma de decisiones, referente al tema de los Incendios Forestales en el estado de Tlaxcala. Posteriormente a la creación del *Data Warehouse, éste* se explotará mediante técnicas adecuadas de minería de datos.

El sistema a desarrollar se fundamenta en Bases de Datos Relacionales y la construcción de cubos OLAP en almacenes de datos. Todo esto es desarrollado en SQL *Server Business Intelligence Development Studio*, para posteriormente ser explotado por Weka (*Waikato Environment for Knowledge Analysis*. Weka es un Entorno para Análisis del Conocimiento de la Universidad de Waikato), es una plataforma de software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL.

1.4 Aportaciones de la investigación

Este proyecto de investigación proporciona una herramienta, de apoyo para la toma de decisiones sobre incendios forestales, mediante la aplicación de técnicas estadísticas de análisis y modelización, encontrando relaciones entre los datos almacenados, es decir, proporciona la capacidad de aprender de los datos del pasado y con ello predecir situaciones futuras en diversos

escenarios. Otra aportación de suma importancia es la optimización tecnológica y económica en entornos del centro de información, estadística y generación de informes.

1.5 Organización de la tesis

El presente trabajo de investigación se encuentra estructurado en 6 capítulos, en los cuales se puede encontrar la siguiente información:

- **Capítulo 1. Introducción:** En este capítulo se plantea el problema a resolver, los objetivos que se plantean alcanzar, las razones que justifican el proyecto, así como el planteamiento de la propuesta que dará solución al problema y los resultados obtenidos. De igual manera se presentan la aportación obtenida producto de la aplicación de la herramienta, después de explotar el Almacén de Datos mediante técnicas de Minería de Datos.

- **Capítulo 2. Estado del Arte:** Se presentan los trabajos que se han realizado en torno a la prevención, detección y extinción de incendios forestales, que tienen como base de desarrollo la creación de *Data Warehouse* y/o explotación mediante técnicas de Minería de Datos.

- **Capítulo 3. Marco teórico:** Este capítulo contiene información sobre los conceptos que soportaron el desarrollo de este proyecto de investigación, así como información básica de las herramientas a utilizar.

- **Capítulo 4. Análisis y diseño:** En este capítulo se recopilan los requerimientos de negocio y se realiza entonces el diseño del modelo del *Data Warehouse* y la descripción de los componentes que lo conforman.

- **Capítulo 5. Implementación y resultados:** Se presentan los resultados obtenidos, al aplicar Técnicas de Minería de datos sobre el *Data Warehouse* antes construido.

- **Capítulo 6. Conclusiones y trabajo a futuro:** Se exponen las conclusiones sobre la funcionalidad y factibilidad del sistema, así como las aportaciones hechas al área y puntos

de vista sobre el posible trabajo futuro que podría mejorar el sistema para hacerlo más robusto.

- Finalmente se presentan las referencias consultadas, para el desarrollo de la tesis.

1.6 Conclusiones

Los incendios forestales tienen un impacto negativo sobre todos los componentes del ecosistema, tanto sobre el medio físico, como biológico y humano. Distintas organizaciones buscan herramientas de apoyo para la toma de decisiones; desde esta prospectiva las tecnologías de *Data Warehouse* y Minería de Datos desempeñan un papel muy importante para la construcción de dichas herramientas de apoyo. En el presente capítulo se presentaron las directrices a seguir para el desarrollo de la herramienta para la toma de decisiones sobre incendios forestales.

2

Estado del arte

En este capítulo se presentan proyectos sobre incendios forestales que han sido realizados bajo el mismo contexto que el de este trabajo de tesis.

2.1 Introducción

Los incendios forestales constituyen una de las causas significativas de la deforestación y la degradación de los ecosistemas, el origen de los problemas generados por los incendios radica fundamentalmente en la irresponsabilidad de algunas personas, ya que el 90% de los incendios forestales ocurridos a nivel mundial, son provocados por el hombre. Los incendios afectan de manera negativa al medio ambiente por la deforestación, la erosión, la pérdida de la biodiversidad y la generación de CO2, los cuales afectan al paisaje y al hábitat de la fauna silvestre [12].

La prevención del fuego es de vital importancia para evitar que se provoquen incendios forestales y/o minimizar sus consecuencias una vez declarados. Bajo este contexto, en los últimos años se ha dado origen a un número importante de trabajos de investigación en el ámbito de los *Data Warehouse* y Minería de Datos.

Al realizar una revisión de los trabajos de investigación que se han desarrollado sobre la prevención de incendios, las principales causas de origen y la extinción de los mismos, se resaltan

los que toman como base de desarrollo la tecnología de *Data Warehouse* y Minería de Datos, para los cuales se presenta un estado del arte detallado.

2.2 Proyectos para la prevención, detección y extinción de incendios forestales

Una gran cantidad de universidades e instituciones forestales, se han dado a la tarea desarrollar herramientas de apoyo para la prevención y extinción de incendios, así como para el estudio de las principales causas que los originan, los cuales se encuentran descritos en la siguiente sección.

2.2.1 Proyecto SIADEX

El sistema SIADEX (sistema inteligente para el diseño asistido de planes de operaciones para la extinción de incendios forestales) es un sistema inteligente capaz de generar, de forma autónoma, planes de extinción de incendios forestales a partir de los datos existentes sobre el terreno (lugar en el que se generó el fuego, entorno, recursos disponibles, etc.) [1].

SIADEX fue creado por un grupo de investigadores de la Escuela Técnica Superior de Ingeniería Informática de la Universidad de Granada (UGR).

SIADEX se ha desarrollado con el producto *IActive Knowledge Studio*, e incorpora el *IActive Decisor* para dotarlo de capacidades de diseño inteligente de procesos. Se encuentra integrado al sistema SIGDIF (Sistema Integrado para la Gestión y Dirección de Incendios Forestales), intercambiando información entre sus subsistemas (El Subsistema de Información Geográfica de la Junta de Andalucía es el encargado de proporcionar cualquier tipo de información territorializada que tenga relación con la problemática de los incendios forestales y de información climática y meteorológica. El subsistema HORUS se encarga del control de posicionamiento y seguimiento de unidades móviles. El subsistema SILVANO se encarga de la gestión del personal que participa en la extinción y de sus turnos de trabajo) y canalizándola a través del sistema INFOGIS (es el programa desde el que acceden los técnicos de extinción y se encarga de coordinar a todos los demás). Para la selección óptima de recursos y la organización del ataque SIADEX, gracias a *IActive Intelligent Decisor*, sigue un proceso de razonamiento basado en los protocolos de actuación estándar definidos por el plan INFOCA (Plan de Prevención y Extinción de Incendios

Forestales en Andalucía) y respeta, de forma sistemática y escrupulosa, la normativa vigente, como la normativa de aviación civil para medios aéreos, los convenios de trabajo para el personal de extinción o los períodos de contratación de los medios externos. Todas las decisiones que propone SIADEX quedan reflejadas en el plan de ataque y son transmitidas selectivamente cada responsable, quienes reciben avisos y alarmas por adelantado de lo que irá ocurriendo.

Igualmente, SIADEX tiene en cuenta todos los factores contextuales como la velocidad del viento (y su posible influencia en el uso de medios aéreos), los cuadrantes de trabajo de los trabajadores y la existencia de varios incendios simultáneamente, etc. En definitiva, gracias a la interconexión con INFOGIS y el resto de subsistemas de SIGDIF, SIADEX disponen de la misma información y conocimiento experto que el director técnico de extinción y puede asistirle en la toma de decisiones de forma óptima.

Estas son algunas de las ventajas que se consiguen gracias a la integración de SIADEX dentro de SIGDIF.

- **Ayuda al técnico de extinción.** Facilita la gestión del conocimiento, información, recursos, etc. a pesar de encontrarse en una situación de estrés y presión como consecuencia de ser un entorno hostil y dinámico.

- **Reducción del tiempo de elaboración del proceso.** *IActive Intelligent Decisor* elabora el proceso de extinción en unos 10 segundos, mientras que los técnicos tardan en elaborarlo de forma manual alrededor de una hora.

- **Homogeniza el proceso de toma de decisiones.** La información se centraliza en un único sistema (Sistema Inteligente), creando un verdadero capital de conocimiento, que antes se encontraba disperso y en distintos soportes, suponiendo un inconveniente a la hora de elaborar el proceso de extinción.

- **Herramienta formativa.** No solo se puede utilizar en un entorno real, sino que también puede ser utilizado para crear episodios pasados, con el fin de ser una herramienta de aprendizaje virtual para técnicos en formación.

- **Seguridad. IActive Intelligent Decisor** es una herramienta sistemática, que explora todas las posibles alternativas teniendo en cuenta todas las restricciones existentes para diseñar un plan de ataque, como legislación vigente, convenios de trabajo, normativas de seguridad, etc. Por tanto, el uso de esta herramienta es una forma de garantizar que los planes obtenidos cumplen todas las normas aplicables de forma automática y demostrable.

En el futuro se prevé dotar a SIADEX de nuevas características que mejoren su funcionamiento dentro de SIGDIF, entre las que cabe destacar el uso de técnicas de aprendizaje automático para perfeccionar su comportamiento y adaptarse a nuevas condiciones que puedan aparecer en futuros incendios [2].

2.2.2 Proyecto detección de incendios forestales a través de imágenes digitales usando árboles de clasificación

El sistema detección de incendios forestales a través de imágenes digitales usando árboles de clasificación, será utilizado por la Secretaría del Medio Ambiente del estado de Puebla (SMRM), puesto que dicha secretaría cuenta con una red de videocámaras que permiten monitorear las regiones del estado propensas a estos fenómenos.

El sistema propone una estrategia que permita procesar las imágenes de incendios forestales y que determine de forma automática de estos eventos en una región. En dicho proyecto se utilizó un método para el procesamiento de las imágenes el cual permita clasificar a nuevas imágenes como dos clases: presencia de incendio forestal o ausencia del mismo.

El desarrollo del proyecto se basa en el proceso KDD (*Knowledge Discovery in Databases*, por sus siglas en inglés), así como en el proceso de análisis de las imágenes basado en particiones, con el fin de medir con mayor precisión las características propias de un incendio forestal. Este enfoque ofrece un resultado confiable para la detección de incendios [3].

2.2.3 Proyecto Análisis de variables sociodemográficas que inciden en incendios forestales, a través de técnicas de *data mining*

El proyecto Análisis de variables sociodemográficas que inciden en incendios forestales, a través de técnicas de *data mining*, es desarrollado en Chile, con el objetivo de encontrar una relación entre las variables sociodemográficas y el inicio del carácter humano de los incendios forestales. Además se persigue encontrar cuáles de estas variables son las más influyentes en un lugar geográfico determinado.

Es de conocimiento público que en Chile, el origen de los incendios forestales es principalmente causado por la acción del hombre, ya sea por negligencia o intencionalidad. Bajo este contexto, nace la idea de analizar las posibles variables sociodemográficas que inciden en el inicio de incendios forestales, variables relacionadas con la organización política, calidad de vida, educación, entre otras, en la población en torno a patrimonios forestales pertenecientes a una gran empresa de la zona sur del país, abarcando parte de la VIII y IX región. Debido a la disponibilidad de grandes bases de datos de los incendios forestales de los últimos 16 años, la herramienta empleada en la investigación es el *data mining*. Herramienta que ofrece una gran versatilidad para el análisis de los datos, donde se utiliza una comparación de distintos algoritmos entre las técnicas de árbol de decisión y redes neuronales. Para el análisis se utilizó el software de libre distribución WEKA.

El problema comienza con la escasa importancia concedida a los factores humanos frente a los físicos en los análisis cuantitativos del riesgo de un incendio forestal. Esta escasa consideración se debe a la dificultad que existe para valorar, modelar y representar espacialmente la influencia humana en el inicio del fuego. Otro aspecto a considerar es que, debido a la dificultad de predecir las singularidades del comportamiento humano, nos encontramos con un alto grado de aleatoriedad a la hora de determinar la ocurrencia de incendios de origen humano. A pesar de ello resulta evidente que los incendios causados por el hombre que se producen en forma reiterada en un determinado ámbito geográfico, no son únicamente reducibles a factores personales individuales y, por tanto, sujetos a reglas del puro azar, sino que este tipo de incendios suelen ser resultado de una pauta social cuyo origen está en las condiciones sociales, sociodemográficas y socioeconómicas de la zona de estudio.

Esta pauta social es la que motiva a esta investigación ir en la búsqueda de nuevas herramientas matemáticas que permitan el entendimiento de las variables sociales que inciden en la ocurrencia de los incendios forestales. La tarea se encuentra en lograr disminuir el número de incendios forestales a través del entendimiento de estas variables, para finalmente construir metodologías útiles en la disminución de los costos asociados a los incendios y lograr una mayor eficiencia en las tareas de prevención [4].

2.2.4 Proyecto Aplicación de un Sistema de Información Geográfica al análisis de los datos de incendios forestales en España

El siguiente proyecto aborda el problema de los incendios forestales en España centrándose en la obtención de información adicional en la toma de decisiones orientadas a la prevención.

El planteamiento de este trabajo responde al interés de generar información de apoyo y complemento, tanto a las actividades de gestión forestal llevadas a cabo por las distintas administraciones, como a otros trabajos realizados por los organismos de investigación (universidades, CSIC, Programa SEXTANTE, Proyecto FIERESTAR, etc.) que generan conocimientos en aspectos relacionados con la prevención y la gestión de los incendios así como en la evaluación y gestión de daños. En este sentido, la metodología que se propone puede facilitar el análisis de la información recogida en los partes de incendios y que se incluye en la base de datos de incendios de la antigua Dirección General de la Biodiversidad (DGB), actualmente Dirección General del Medio Natural y Política Forestal.

La metodología de trabajo utilizada consta de dos etapas fundamentales, una fase de preparación en la que partiendo de unos datos iníciales se seleccionan, limpian y transforman preparándolos para la fase de análisis en la que se aplican distintos procedimientos para extraer nueva información o información resumida.

Una vez preparados los datos se aborda la fase de análisis de datos utilizando las herramientas de análisis espacial suministradas por GeoMedia, el Sistema de Información Geográfica (SIG) empleado. GeoMedia permite capturar, relacionar, manipular, analizar y mostrar información referenciada geográficamente. Por tanto la información puede ser visualizada en mapas temáticos, es decir se pueden representar fenómenos geográficos, cualitativos y cuantitativos, sobre una base cartográfica simplificada.

En el caso de este proyecto la información se ha analizado por años incorporando de esta forma la dimensión temporal al análisis espacial. Asimismo los mapas temáticos obtenidos en este proyecto se presentan con dos niveles de resolución espacial: a) la provincia y b) la cuadrícula de 10 km * 10 km.

a) Con resolución espacial de provincia se elaboraron los mapas temáticos que emplean las bases de datos de la DGB y de GIS y que se relacionan a través de un campo común que contiene el codeo dado a las provincias.
b) Cuando la unidad espacial es una cuadrícula de 10 km * 10 km, que divide espacialmente toda la superficie de España.

Este estudio se amplió con la información espacial contenida en la base de datos del *Corine Land Cover 2000* (CLC2000), que refleja los diferentes usos de suelo, creando los mapas temáticos que contenían el número total de incendios según el uso de suelo por cuadrícula de 10 km *10 km por zona y año, para el periodo 1995-2004.

Por último se crearon mapas temáticos con la resolución 10 km * 10 km a partir de la base de datos del Segundo Inventario Forestal Nacional (2° IFN) que contiene información sobre los tipos de propiedad de los montes. En este caso se analizó el número total de incendios según el tipo de propiedad por cuadrícula de 10 km * 10 km y zona y año. Para el periodo 1995-2004 [5].

2.3 Incendios forestales y la toma de decisiones

Los sistemas antes mencionados son de gran utilidad para la prevención, detección y extensión de incendios. Sin embargo cada uno de ellos resuelve problemas en específico, por ejemplo: SIADEX genera planes para extinción de incendios, otro sistema identifica la zona en la que existe un incendio, por medio del análisis de imágenes. También se presentó un sistema que estudia las variables socio demográficas que originan un incendio.
Con el desarrollo del "Sistema para la Toma de Decisiones para Incendios Forestales del Estado de Tlaxcala" se creó un almacén de datos, en el cual se recopiló la información existente, sobre los incendios forestales que se han generado en el estado, para posteriormente analizarlo mediante operaciones OLAP y minería de datos.

2.4 Conclusiones

En este capítulo se presentaron diferentes proyectos relacionados con la detección, prevención y extinción de incendios, que han sido desarrollados en varios países teniendo como base de desarrollo la edificación de *Data Warehouse* y el análisis mediante minería de datos.

3

Marco teórico

El presente capítulo ofrece una breve descripción de los conceptos teóricos sobre Data Warehouse (DW), minería de datos y herramientas que serán empleadas en el desarrollo, así como algoritmos relacionados con el trabajo de tesis.

3.1 Almacén de datos

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos digitales y otras fuentes ha crecido espectacularmente en las últimas décadas. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Aparte de su función de "memoria de la organización", la información histórica es útil para explicar el pasado, entender el presente y predecir la información futura. La mayoría de las decisiones de empresas, organizaciones e instituciones se basan en información sobre experiencias pasadas extraídas de fuentes muy diversas. Además, ya que los datos pueden proceder de fuentes diversas y pertenecer a diferentes dominios, parece clara la inminente necesidad de analizar los mismos para la obtención de información útil para la organización [6].

Según definió *Bill Inmon* [7], el *Data Warehouse* se caracteriza por ser:

Integrado: Los datos almacenados en el *Data Warehouse* deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

Temático: Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

Histórico: El tiempo es parte implícita de la información contenida en un *Data Warehouse*. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el *Data Warehouse* sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el *Data Warehouse* se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

No volátil: El almacén de información de un *Data Warehouse* existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del *Data Warehouse* la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

3.1.1 OLTP y OLAP

Existen dos tipos de procesamientos orientados al análisis de datos, los cuales corresponden a las siglas OLTP y OLAP:

- **OLTP:** Constituye el trabajo primario en un sistema de información. Consiste en realizar transacciones, es decir, actualizaciones y consultas a la base de datos con un objetivo operacional: hacer funcionar las aplicaciones de la organización, proporcionar información sobre el estado del sistema de información y permitir actualizarlo conforme va variando la realidad del contexto de la organización.

- **OLAP (***On-Line Analytical Processing***):** Engloba un conjunto de operaciones, exclusivamente de consulta, en las que se requiere agregar y cruzar gran cantidad de información. El objetivo es realizar informes y resúmenes, generalmente para el apoyo en la toma de decisiones.

Una característica de ambos procesamientos es que se pretende que sean "on-line", es decir, que sean relativamente "instantáneos" y se puedan realizar en cualquier momento (en tiempo real). Esto parece evidente e imprescindible para el OLTP, pero no esta tan claro que esto sea posible para algunas consultas muy complejas realizadas por el OLAP.

Hasta hace pocos años, y todavía existente en muchas organizaciones y empresas, ambos tipos de procesamiento (OLTP y OLAP) se realizaran sobre la misma base de datos transaccional, con lo cual se generan dos problemas fundamentales:

- Las consultas OLAP perturban el trabajo transaccional diario de los sistemas de información originales. Al ser consultas complejas y que involucran muchas tablas y agrupaciones, suelen consumir gran parte de los recursos del sistema de gestión de base de datos. El resultado es que durante la ejecución de estas consultas, las operaciones OLTP, se resienten: las aplicaciones van más lentas, las actualizaciones consumen mucho tiempo y el sistema puede incluso llegar a colapsarse. De este hecho viene el nombre familiar que se les da a las consultas OLAP: *"killer queries"* (consultas asesinas). Como consecuencia, muchas de estas consultas se deben realizar por la noche o en fines de semana, con lo que en realidad dejan de ser "*on-line".*

- La base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Esto significa que, aunque tuviéramos el sistema dedicado exclusivamente para realizar una consulta OLAP, dicha consulta puede requerir mucho tiempo, pero no solo por ser compleja intrínsecamente, sino porque el esquema de la base de datos no es el más adecuado para este tipo de consultas.

Ambos problemas implican que va a ser prácticamente imposible (a un costo de hardware razonable, lógicamente) realizar un análisis complejo de la información en tiempo real si ambos procesamientos se realizan sobre la misma base de datos.

Desde esta perspectiva, se separa definitivamente la base de datos con fines transaccionales de la base de datos con fines analíticos y es así que nacen los almacenes de datos [6].

3.1.2 Almacenes de datos y bases de datos transaccionales

Un almacén de datos es un conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas.

La ventaja fundamental de un almacén de datos es su diseño específico y su separación de la base de datos transaccional. Un almacén de datos:

- Facilita el análisis de los datos en tiempo real (OLAP).
- No disturba el OLTP de las bases de datos originales.

Las diferencias mostradas en la tabla 3.1 se distingue la manera de estructurar y diseñar almacenes de datos respecto a la forma tradicional de hacerlo con bases de datos transaccionales.

Tabla 3. 1 *Diferencia entre la base de datos transaccional y el almacén de datos.*

	BASE DE DATOS TRANSACCIONAL	**ALMACÉN DE DATOS**
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Tipo de datos	Datos de funcionamiento de la organización.	Datos útiles para el análisis, la sumarización, etc.
Características de los datos	Datos de funcionamiento, cambiantes, internos, incompletos...	Datos históricos, datos internos y externos, datos descriptivos...
Modelos de datos	Relacional, relacional-extendido Datos normalizados.	Multidimensionales Esquema estrella, copo de nieve. Parcialmente desnormalizados,
Número y tipo de usuarios	Cientos/miles: aplicaciones, operarios, administrador de la base de datos.	Decenas: directores, ejecutivos, analistas (granjeros, mineros).
Acceso	SQL. Lectura y escritura.	SQL y herramientas propias *(slice & dice, drill, roll, pivot..).* Lectura.

Aunque ambas fuentes de datos (transaccional y almacén de datos) están separadas, es importante destacar que gran parte de los datos que se incorporan en un almacén de datos provienen de la base de datos transaccional. Esto supone desarrollar una tecnología de volcado y mantenimiento de datos desde la base de datos transaccional al almacén de datos. Además, el almacén de datos debe integrar datos externos, con lo que en realidad debe estar actualizándose frecuentemente de diferentes fuentes. El almacén de datos pasa a ser un integrador o recopilador de información de diferentes fuentes [6], ver figura 3.1.

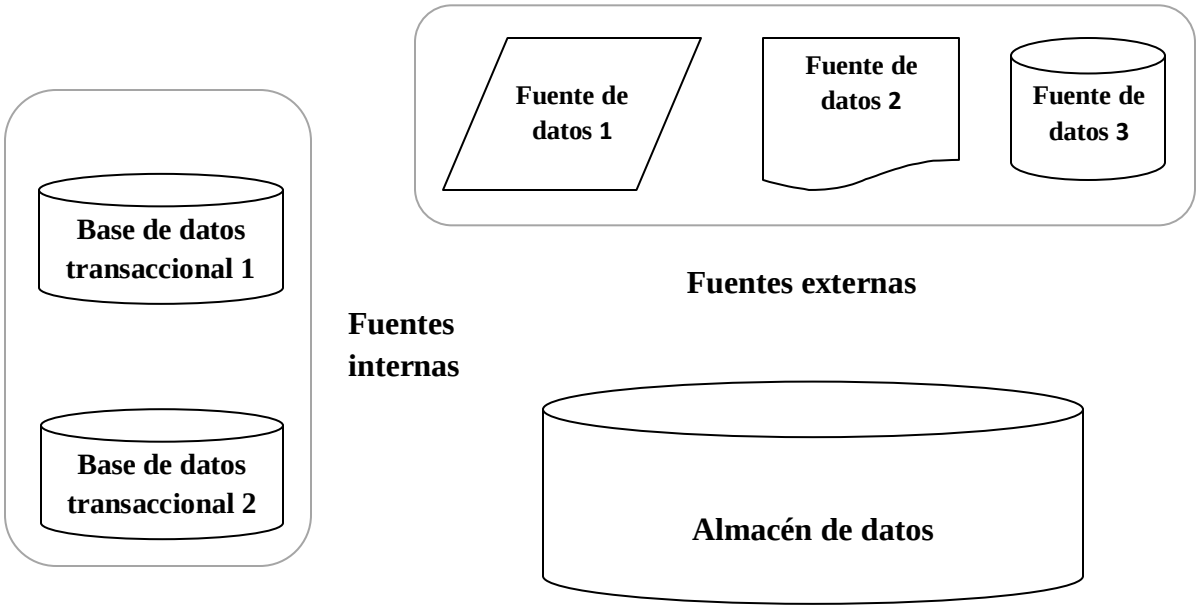

Figura 3. 1 *El almacén de datos como integrador de diferentes fuentes de datos* [6].

3.1.3 Arquitectura de los almacenes de datos

Un almacén de datos recoge, fundamentalmente, datos históricos, es decir, hechos, sobre él contexto en el que se desenvuelve la organización. Los hechos son, por tanto, el aspecto central de los almacenes de datos.

3.1.4 Modelo multidimensional

Los almacenes de datos y herramientas OLAP se basan en un modelo de datos multidimensional. Este modelo visualiza los datos en forma de un cubo de datos [8]. Los datos se organizan en torno a los hechos, que tienen atributos o medidas que pueden verse en mayor o menor detalle según ciertas dimensiones [6].

En términos generales, las dimensiones son las perspectivas o entidades respecto de las cuales una organización quiere llevar registros. Por ejemplo, una tienda puede crear un almacén de datos de ventas, a fin de mantener los registros de ventas de la tienda con respecto a las dimensiones tiempo, artículo, sucursal, y ubicación. Cada dimensión puede tener una tabla asociada a ella, llamada tabla de dimensiones. Por ejemplo, una tabla de dimensiones para artículo puede contener los atributos nombre de artículo, marca y tipo. Las tablas de dimensiones pueden ser especificadas por los usuarios o expertos, o se genera automáticamente y se ajustan en base a distribución de datos [8].
Las dimensiones por lo general responden a las preguntas ¿Cuándo?, ¿Qué? y ¿Dónde? [6].

Un modelo de datos multidimensional se organiza normalmente en torno a un tema central, como las ventas, por ejemplo. Este tema está representado por una tabla de hechos. Los hechos son medidas numéricas. Piense en ellos como las cantidades por las que desea analizar las relaciones entre las dimensiones. Ejemplos de hechos para un almacén de datos de ventas son: pesos vendidos (importe de las ventas en pesos), las unidades vendidas (número de unidades vendidas), y la cantidad presupuestada. La tabla de hechos contiene los nombres de los hechos o medidas, así como claves para cada una de las tablas de dimensiones relacionadas [8].
Las medidas responden generalmente a la pregunta ¿Cuánto? [6].

Aunque solemos pensar en los cubos 3-D como estructuras geométricas, en el almacenamiento de los datos el cubo de datos es n-dimensional. El cubo de datos es una metáfora para el almacenamiento de datos multidimensional. El almacenamiento físico real de estos datos puede diferir de la representación lógica. Lo importante a recordar es que los cubos de datos son n-dimensional y no limitarse a los datos en 3-D.

En la literatura de investigación de almacenamiento de datos, un cubo de datos, se refiere a menudo como un paralelepípedo. Dado un conjunto de dimensiones, podemos generar un

paralelepípedo de cada uno de los posibles subconjuntos de las dimensiones. El resultado sería un conjunto de paralelepípedos, cada uno muestra los datos en un nivel diferente de resumen, o grupo. El paralelepípedo que tiene el nivel más bajo de resumen se llama paralelepípedo de base.

El modelo de datos más popular para un almacén de datos es un modelo multidimensional. Este modelo puede existir en forma de un esquema de estrella, un esquema de copo de nieve, o esquema de constelación de hechos.

Esquema en estrella: El paradigma de modelado más común es el esquema en estrella, en la que el almacén de datos contiene una gran tabla central (tabla de hechos) que contiene la mayor parte de los datos, sin redundancia, y un conjunto de pequeños asistente tablas (tablas de dimensiones), una para cada dimensión. El gráfico de la figura 3.2 se asemeja a una estrella, con las tablas de dimensiones alrededor de la tabla de hechos.

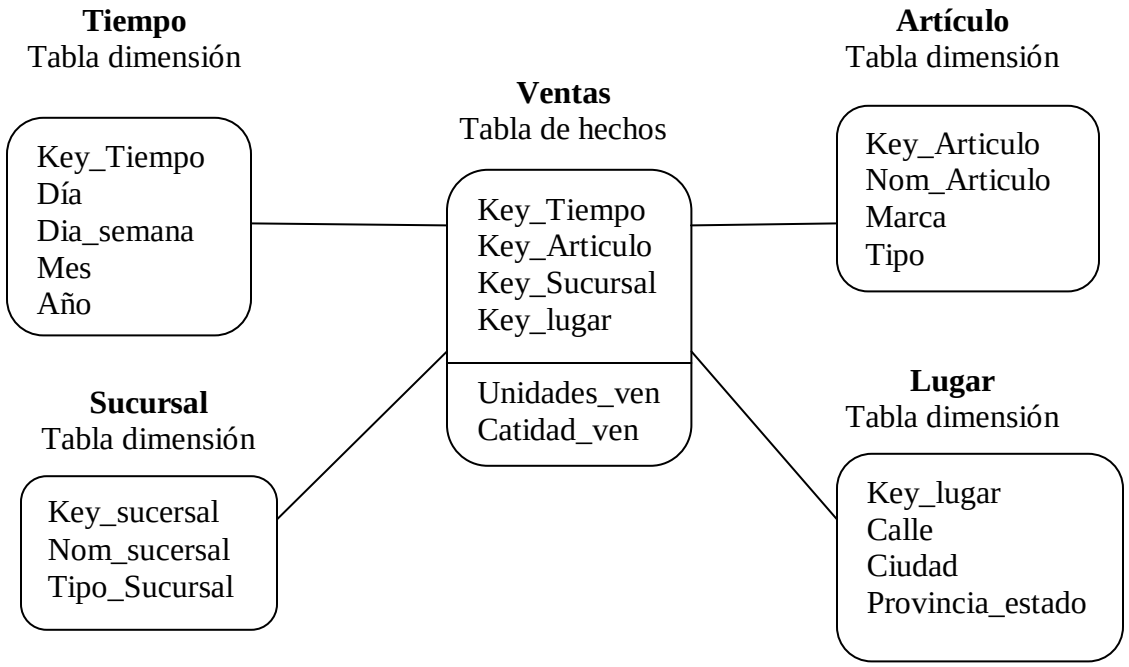

Figura 3. 2 *Esquema de estrella del almacén de datos para ventas.*

Esquema de Copo de nieve: El esquema de copo de nieve es una variante del modelo de estrella, donde algunas tablas de dimensiones se normalizan, lo que divide los datos en tablas adicionales. La figura 3.3 constituye una forma similar a un copo de nieve.

La diferencia principal entre los modelos copo de nieve y de estrella es que las tablas de dimensiones del modelo de copo de nieve se pueden mantener en forma normalizada para reducir la redundancia. Dicha tabla es fácil de mantener y ahorra espacio de almacenamiento. Sin embargo, este ahorro de espacio es insignificante en comparación con la magnitud típica de la tabla de hechos. Además, la estructura del copo de nieve puede reducir la eficacia de la navegación, ya que es necesario realizar un *join* en la ejecución de una consulta. En consecuencia, el rendimiento del sistema puede verse afectado. Por lo tanto, aunque el esquema de copo de nieve reduce la redundancia, no es tan popular como el esquema de estrella en el diseño de almacenamiento de datos.

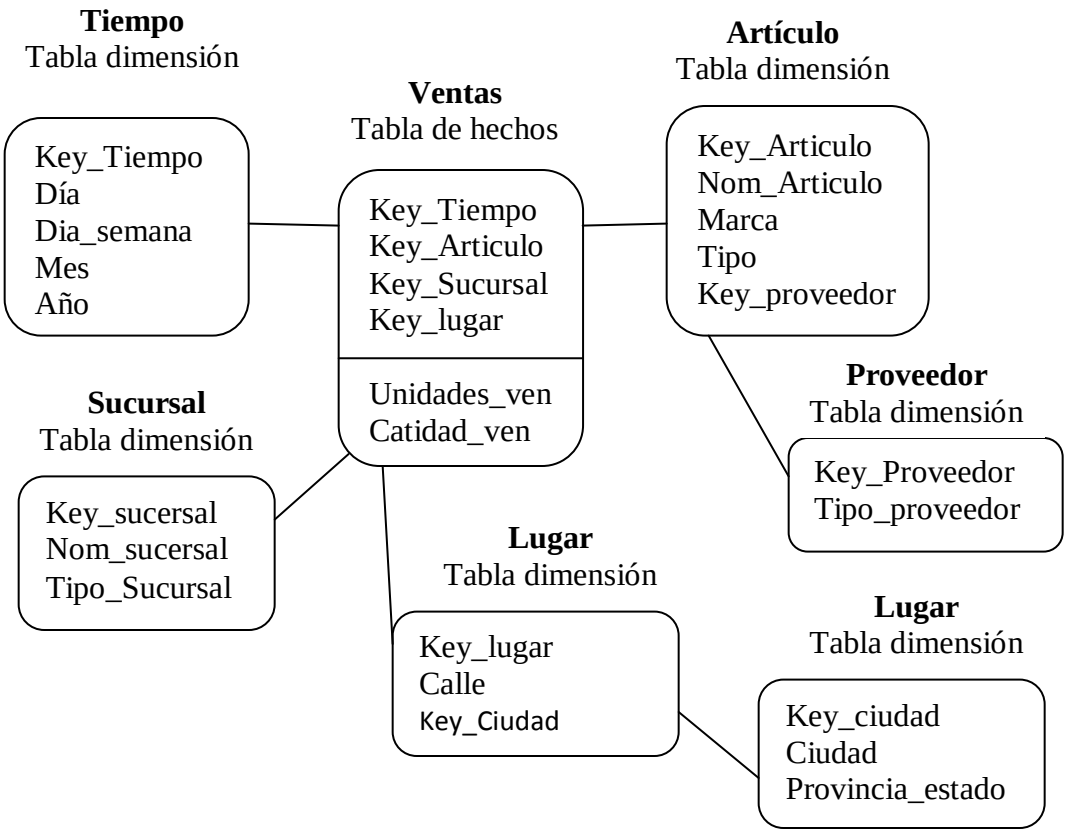

Figura 3. 3 *Esquema copo de nieve del almacén de datos para ventas.*

Constelación de hechos: Aplicaciones sofisticadas pueden requerir varias tablas de hechos para compartir tablas de dimensiones. Este tipo de esquema puede ser visto como una colección de estrellas, y por lo tanto, se denomina esquema galaxia o una constelación de hechos [8], ver figura 3.4.

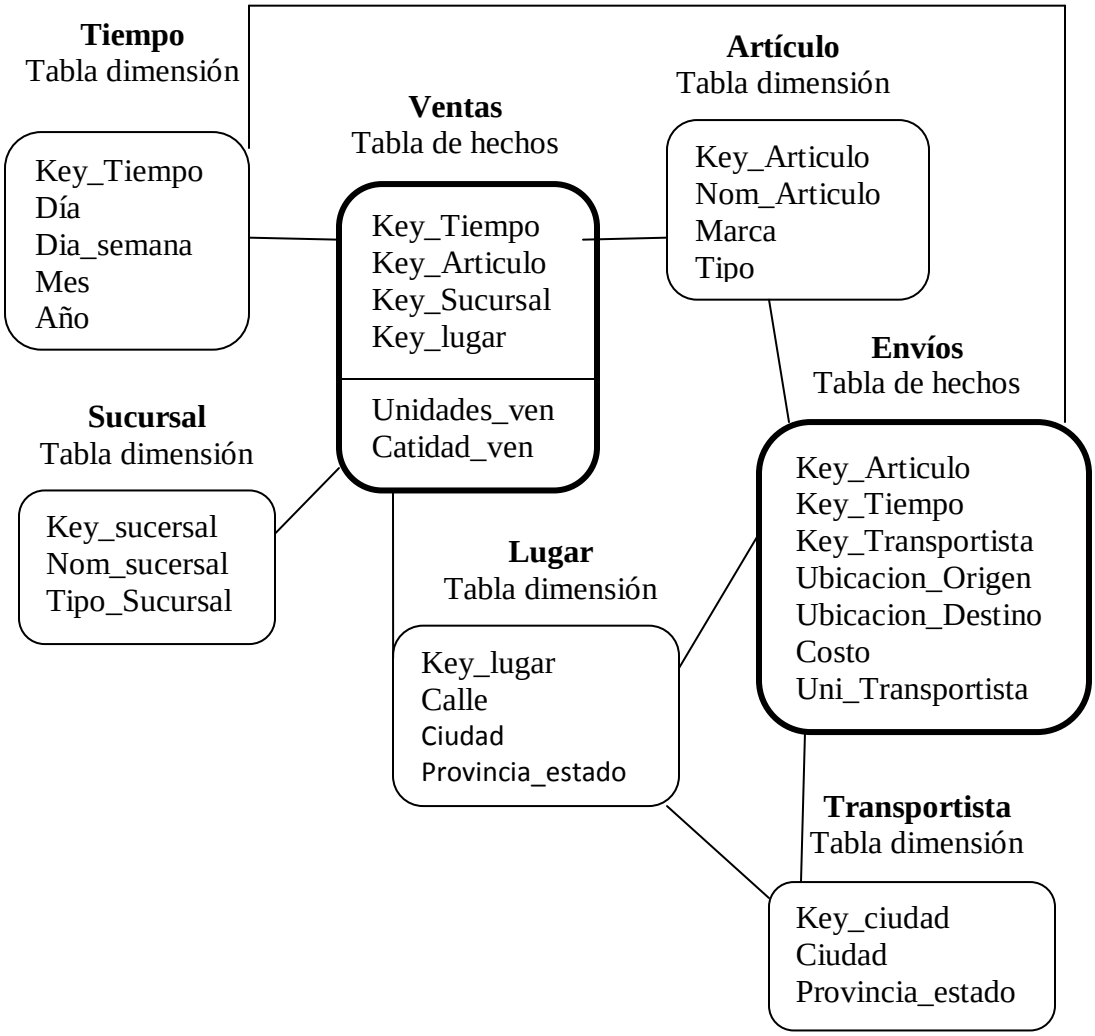

Figura 3. 4 *Esquema constelación de hechos de un almacén de datos para ventas y envíos.*

3.1.5 *Data Warehouse vs. Datamart*

En un contexto de *Data Warehouse*, el término duplicación se refiere a la creación de *Data Marts* locales o departamentales basados en subconjuntos de la información contenida en el *Data Warehouse* central o maestro.

Según define *Meta Group*, "un *Data Mart* es una aplicación de *Data Warehouse*, construida rápidamente para soportar una línea de negocio simple". Los *Data Marts*, tienen las mismas características de integración, no volatilidad, orientación temática y no volatilidad que el *Data Warehouse*. Representan una estrategia de "divide y vencerás" para ámbitos muy genéricos de un *Data Warehouse*.

Esta estrategia es particularmente apropiada cuando el *Data Warehouse* central crece muy rápidamente y los distintos departamentos requieren sólo una pequeña porción de los datos contenidos en él. La creación de estos *Data Marts* requieren algo más que una simple réplica de los datos: se necesitarán tanto la segmentación como algunos métodos adicionales de consolidación. La primera aproximación a una arquitectura descentralizada de *Data Mart*, podría ser originada de una situación como la descrita en la figura 3.5.

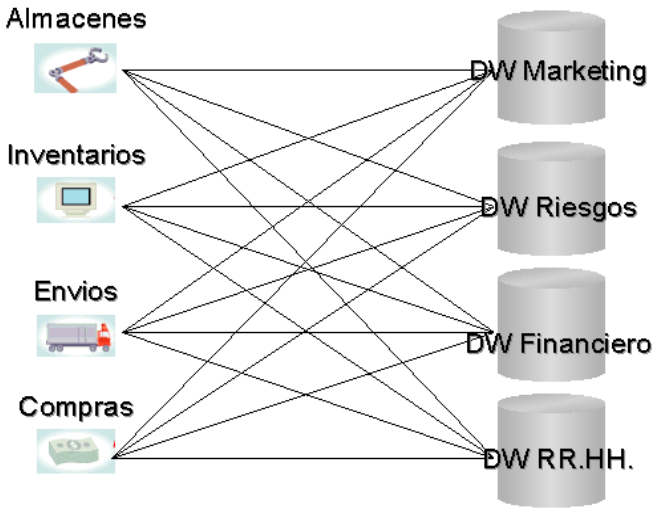

Figura 3. 5 *Aproximación a una arquitectura descentralizada de Data Mart* [14].

El departamento de *Marketing*, emprende el primer proyecto de *Data Warehouse* como una solución departamental, creando el primer *Data Mart* de la empresa. Visto el éxito del proyecto, otros departamentos, como el de Riesgos, o el Financiero se lanzan a crear sus *Data Marts*. *Marketing*, comienza a usar otros datos que también usan los *Data Marts* de Riesgos y Financiero, y estos hacen lo propio. Esto parece ser una decisión normal, puesto que las necesidades de información de todos los *Data Marts* crecen conforme el tiempo avanza. Cuando esta situación evoluciona, el esquema general de integración entre los *Data Marts* pasa a ser, la de la figura 3.6.

En esta situación, es fácil observar cómo este esquema de integración de información de los *Data Marts*, pasa a convertirse en un rompecabezas en el que la gestión se ha complicado. No obstante, lo que ha fallado no es la integración de *Data Marts*, sino su forma de integración.

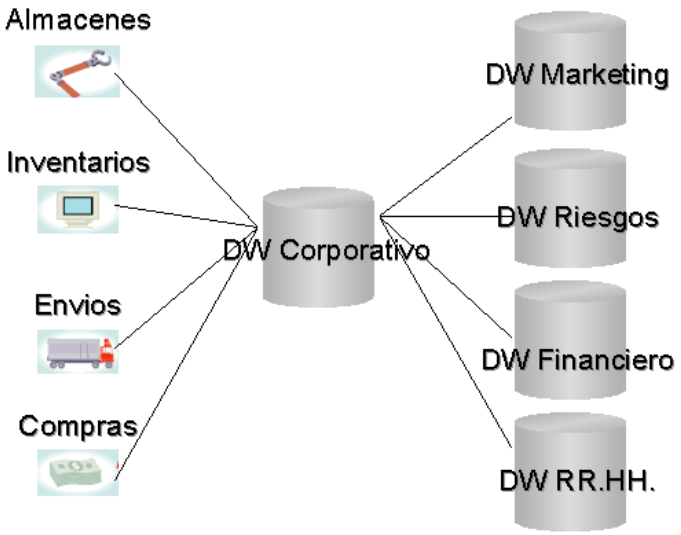

Figura 3. 6 *Integración entre los Data Marts* [14].

En efecto, un enfoque más adecuado sería la coordinación de la gestión de información de todos los *Data Marts* en un *Data Warehouse* centralizado. En esta situación los *Data Marts* obtendrían la información necesaria, ya previamente cargada y depurada en el *Data Warehouse* corporativo, simplificando el crecimiento de una base de conocimientos a nivel de toda la empresa. Esta simplificación provendría de la centralización de las labores de gestión de los *Data Marts*, en el *Data Warehouse* corporativo, generando economías de escala en la gestión de los *Data Marts* implicados [14].

3.1.6 Explotación de un almacén de datos

En realidad, un modelo de datos se compone de estructuras y operadores sobre dichas estructuras. El modelo multidimensional se basa en un conjunto de *data marts*, que, generalmente, son estructuras de datos en estrella jerárquica.

Para completar el modelo multidimensional debemos definir una serie de operadores sobre la estructura. Los operadores más importantes asociados a este modelo son:

- ***Drill***: Se trata de disgregar los datos (mayor nivel de detalle o desglose, menos sumarización) siguiendo los caminos de una o más dimensiones.
- ***Roll:*** Se trata de agregar los datos (menor nivel de detalle o desglose, más sumarización o consolidación) siguiendo los caminos de una o más dimensiones.
- ***Slice & Dice:*** se seleccionan y se proyectan datos, de acuerdo a los operadores del álgebra relacional.
- ***Pivot:*** se reorientan las dimensiones.

Normalmente, estos operadores se llaman operadores OLAP, operadores de análisis de datos u operadores de almacén de datos. Los operadores básicos permiten realizar las mismas consultas de proyección, selección y agrupamiento que se pueden hacer en SQL.
Lo interesante de los operadores *drill, roll, slice & dice y pivot*, es que permiten modificar la consulta realizada, sin necesidad de realizar otra. En realidad son "navegadores" de informes, más que operadores por sí mismos [6].

3.1.7 Implementación del almacén de datos. Diseño

Una de las razones para crear un almacén de datos separado de la base de datos operacional era conseguir que el análisis se pudiera realizar de una manera eficiente. Con el objetivo de obtener la eficiencia deseada, los sistemas de almacenes de bases de datos pueden implementarse utilizando dos tipos de esquemas físicos:

- ROLAP *(Relational OLAP):* Físicamente, el almacén de datos se construye sobre una base de datos relacional.

- MOLAP *(Multidimensional OLAP):* Físicamente, el almacén de datos se construye sobre estructuras basadas en matrices multidimensionales.

Las ventajas del ROLAP son, en primer lugar, que se pueden utilizar directamente sistemas de gestión de bases de datos genéricos y herramientas asociadas: SQL, restricciones, disparadores, etc. En segundo lugar, la formación y el costo necesario para su implementación es generalmente menor. Las ventajas del MOLAP son su especialización, la correspondencia entre el nivel lógico y el nivel físico. Esto hace que el MOLAP sea generalmente más eficiente.

Como se ha mencionado, la ventaja de los ROLAP es que pueden utilizar tecnología y nomenclatura de los sistemas de bases de datos relacionales. Esto tiene el riesgo de que en algunos casos se pueda decidir mantener parte de la base de datos transaccional o inspirarse en su organización (manteniendo claves ajenas, claves primarias, conservando parte de la normalización, etc.).

Una de las maneras más eficientes de implementar un *datamart* multidimensional mediante bases de datos relacionales se basa en ignorar casi completamente la estructura de los datos en las fuentes de origen y utiliza una estructura nueva denominada *starflake*. Esta estructura combina los esquemas en estrella (*star)* y en estrella jerárquica o copo de nieve (*snowflake)*.

Este diseño proporciona la realización de consultas OLAP de una manera eficiente, así como la aplicación de los operadores específicos:

- Las tablas copo de nieve permiten realizar vistas o informes utilizando diferentes grados de detalle sobre varias dimensiones. Al estar normalizadas permiten seleccionar datos dimensionales de manera no redundante. Esto es especialmente útil para los operadores *drill, slice & dice y pivot*.
- Las tablas estrella son, como hemos dicho, tablas de apoyo, que representan "pre-concatenaciones" o "pre-junciones" (pre-j*oins*) entre las tablas copo de nieve. El propósito de las tablas estrella es evitar concatenaciones costosas cuando se realizan operaciones de *roll-up*.

En cambio los sistemas MOLAP tienen algunos inconvenientes:

- Se necesitan sistemas específicos. Esto supone un costo de *software* mayor y generalmente compromete la portabilidad, al no existir estándares sobre MOLAP tan extendidos como los estándares del modelo relacional.
- Al existir un gran acoplamiento entre la visión externa y la implementación, los cambios en el diseño del almacén de datos obligan a una reestructuración profunda del esquema físico y viceversa.
- Existe más des-normalización que en las ROLAP. En muchos casos un almacén de datos MOLAP ocupa más espacio que su correspondiente ROLAP.

Quizá la parte de diseño de almacenes de datos es una de las áreas más abiertas y donde existe menos convergencia. Las razones son múltiples pero, fundamentalmente, se resumen en que los almacenes de datos se han originado principalmente desde el ámbito industrial y no académico, que el fin inicial del almacén de datos era realizar OLAP eficiente, con lo que el énfasis recaía fundamentalmente en los niveles lógico y físico. A pesar de todo esto, podemos identificar cuatro pasos principales a la hora de diseñar un almacén de datos (en realidad estos pasos se han de seguir para cada *datamart*):

1. Elegir para modelar un "proceso" o "dominio" de la organización sobre el que se deseen realizar informes complejos frecuentemente, análisis o minería de datos.

2. Decidir el hecho central y el "gránulo" (nivel de detalle) máximo que se va a necesitar sobre él. En general, siempre hay que considerar gránulos finos por si más adelante se fueran a necesitar, a no ser que haya restricciones de tamaño importantes. Precisamente, el almacén de datos se realiza, entre otras cosas, para poder agregar eficientemente, por lo que un almacén de datos demasiado detallado no compromete, en principio, la eficacia.

3. Identificar las dimensiones que caracterizan el "dominio" y su grafo o jerarquía de agregación, así como los atributos básicos de cada nivel. No se deben incluir atributos descriptivos más que lo imprescindible para ayudar en la visualización. En cambio atributos informativos del estilo "es festivo", "es fin de semana", "es festival", etc., son especialmente interesantes de cara a agregaciones y selecciones que detecten patrones. Las dimensiones varían mucho de un dominio a otro, aunque respondan a preguntas como "que", "quien",

"donde", "de donde", "cuando", "como", etc. El tiempo siempre es una (o más de una) de las dimensiones presentes.

4. Determinar y refinar las medidas y atributos necesarios para los hechos y las dimensiones. Generalmente las medidas de los hechos son valores numéricos agregables (totales, cuentas, medias...) y suelen responder a la pregunta "cuanto".

3.1.8 Extracción, transformación y carga del almacén de datos

Finalmente, si se ha decidido diseñar un almacén de datos, y ya está implementado mediante tecnología ROLAP o MOLAP, el siguiente paso es extraer, transformar y cargar los datos.

En realidad, la extracción, transformación y carga de un almacén de datos es uno de los aspectos que más esfuerzo requiere (alrededor de la mitad del esfuerzo necesario para implantar un almacén de datos), y, de hecho, suele existir un sistema especializado para realizar estas tareas, denominado sistema ETL *(Extraction, Transformation, Load).* Dicho sistema no se compra en el supermercado ni se descarga de Internet, sino que:

- La construcción del ETL es responsabilidad del equipo de desarrollo del almacén de datos y se realiza específicamente para cada almacén de datos.

El sistema ETL se encarga de realizar muchas tareas:

- **Lectura de datos transaccionales:** se trata generalmente de obtener los datos mediante consultas SQL sobre la base de datos transaccional. Generalmente se intenta que esta lectura sea en horarios de poca carga transaccional (fines de semana o noches). Para la primera carga los datos pueden encontrarse en históricos y es posible que en distintos formatos. Este hecho condiciona muchas veces el número de años que se puede incluir en el almacén de datos.

- **Incorporación de datos externos:** generalmente aquí se deben incorporar otro tipo de herramientas, como *wrappers*, para convertir texto, hojas de cálculo o HTML en XML o en tablas de base de datos que se puedan integrar en el almacén de datos.

- **Creación de claves:** en general se recomienda crear claves primarias nuevas para todas las tablas que se vayan creando en el almacenamiento intermedio o en el almacén de datos.

- **Integración de datos:** consiste en muchos casos en la fusión de datos de distintas fuentes, detectar cuando representan los mismos objetos y generar las referencias y restricciones adecuadas para conectar la información y proporcionar integridad referencial.
- **Obtención de agregaciones:** si se sabe que cierto nivel de detalle no es necesario en ningún caso, una primera fase de agregación se puede realizar aquí.

- **Limpieza y transformación de datos:** Parte de la limpieza y la transformación necesaria para organizar el almacén se realiza por el ETL. Se trata, como veremos, de evitar datos redundantes, inconsistentes, estandarizar medidas, formatos, fechas, tratar valores nulos, etc.

- **Creación y mantenimiento de metadatos:** para que todo el ETL pueda funcionar es necesario crear y mantener metadatos sobre el propio proceso ETL y los pasos realizados y por realizar.

- **Identificación de cambios:** esto se puede realizar de muy distintas maneras: mediante una carga total cada vez que haya un cambio, mediante comparación de instancias (uso de archivos delta), mediante marcas de tiempo *(time stamping)* en los registros, mediante disparadores, mediante el archivo de *log* o mediante técnicas mixtas. Algunas son muy ineficientes (carga total o uso de disparadores) y otras son muy complejas de implementar (archivo de *log)*. Generalmente, por tanto, se utilizan técnicas mixtas.

- **Planificación de la carga y mantenimiento:** consiste en definir las fases de carga, el orden, para evitar violar restricciones de integridad, del mismo modo que se realizan las migraciones, y las ventanas de carga, con el objetivo de poder hacer la carga sin saturar ni la base de datos transaccional, así como el mantenimiento sin paralizar el almacén de datos.

- **Indexación:** finalmente se han de crear índices sobre las claves y atributos del almacén de datos que se consideren relevantes (niveles de dimensiones, tablas de hechos, etc.).

- **Pruebas de calidad:** en realidad se trata de definir métricas de calidad de datos del almacén de datos, así como implantar un programa de calidad de datos, con un responsable de calidad que realice un seguimiento, especialmente si el almacén de datos se desea utilizar para el apoyo en decisiones estratégicas o especialmente sensibles.

Generalmente, para realizar todas estas tareas, los sistemas ETL se basan en un repositorio de datos intermedio, como se muestra en la Figura 3.7. Esto puede parecer que ya es abusar de recursos, al tener además de la base de datos transaccional y el almacén de datos un tercer repositorio de datos de similar magnitud. Sin embargo, este almacenamiento intermedio es extremadamente útil, ya que hay tareas que no se pueden realizar en el sistema transaccional ni en el almacén de datos.

Con ello, muchos procesos del ETL, incluidos el mantenimiento, se pueden realizar en gran medida sin paralizar ni la base de datos transaccional ni el almacén de datos.

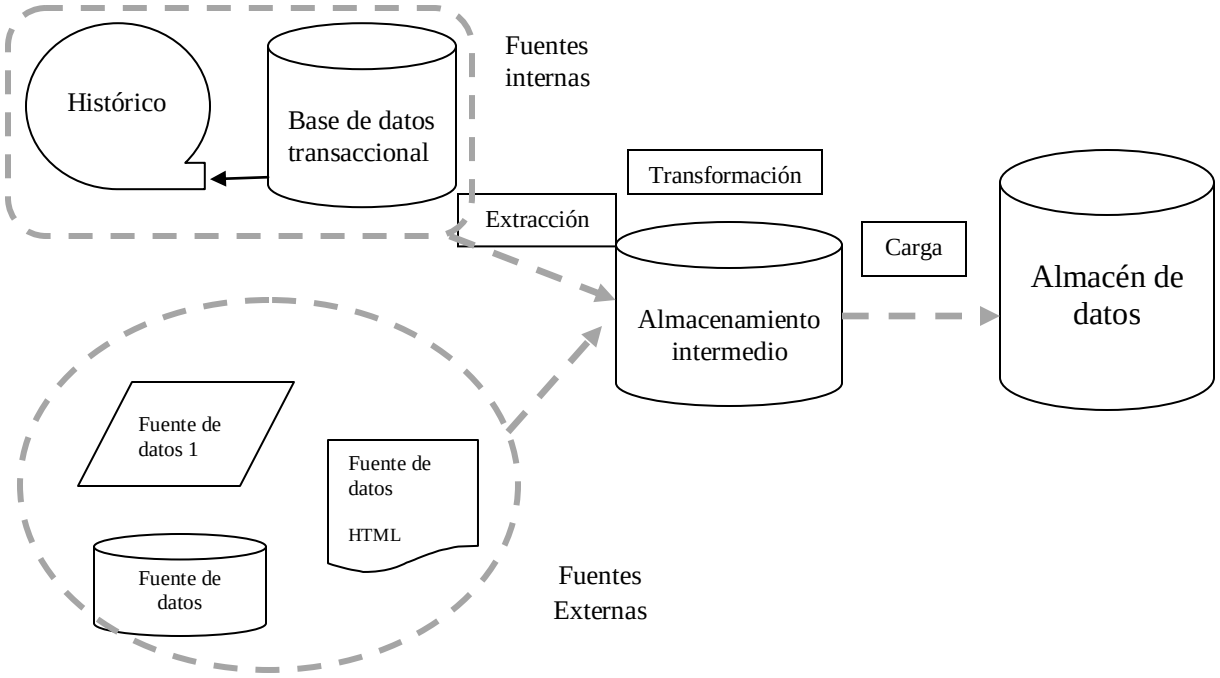

Figura 3. 7 *El sistema ETL basado en un repositorio intermedio* [6].

3.1.9 Almacenes de datos y minería de datos

El concepto de almacenes de datos nace hace más de una década ligado al concepto de EIS *(Executive Information System)*, el sistema de información ejecutivo de una organización. En realidad, cuando están cubiertas las necesidades operacionales de las organizaciones se plantean herramientas informáticas para asistir o cubrir en las necesidades estratégicas.

La definición original de almacén de datos es una "colección de datos, orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar en las decisiones de dirección". Con la difusión cada vez mayor de las herramientas de *business intelligence y* OLAP, podríamos pensar que los almacenes de datos no se aplican en otros ámbitos: científicos, médicos, ingenieriles, académicos, donde no se tratan con las variables y problemáticas típicas de las organizaciones y empresas.

Al contrario, en realidad, los almacenes de datos pueden utilizarse de muy diferentes maneras, y pueden agilizar muchos procesos diferentes de análisis. A un almacén de datos se le pueden dar diferentes aplicaciones y usos: herramientas de consultas e informes, herramientas EIS, herramientas OLAP y herramientas de minería de datos.

Según el carácter de los usuarios se les puede catalogar en dos grandes grupos:

- **"picapedreros" (o "granjeros"):** se dedican fundamentalmente a realizar informes periódicos, ver la evolución de indicadores, controlar valores anómalos, etc.
- **"exploradores":** encargados de encontrar nuevos patrones significativos utilizando técnicas OLAP o de minería de datos. La estructura del almacén de datos y sus operadores facilita la obtención de diferentes vistas de análisis o "vistas minables".

Esta diferencia, y el hecho de que se catalogue como "exploradores" a aquellos que utilizan técnicas OLAP o minería de datos, no nos debe hacer confundir las grandes diferencias de un análisis clásico, básicamente basado en la agregación, la visualización y las técnicas descriptivas estadísticas con un uso genuino de minería de datos que no transforma los datos en otros datos (más o menos agregados) sino que transforma los datos en conocimiento (o mas humildemente, en reglas o modelos). Un aspecto a destacar es que el nivel de agregación para los

requerimientos de análisis OLAP puede ser mucho más grueso que el necesario para la minería de datos.

Los almacenes de datos no son imprescindibles para hacer extracción de conocimiento a partir de datos. En realidad, se puede hacer minería de datos sobre un simple archivo de datos. Sin embargo, las ventajas de organizar un almacén de datos se amortizan sobradamente a medio y largo plazo. Esto es especialmente patente cuando nos enfrentamos a grandes volúmenes de datos, o estos aumentan con el tiempo, o provienen de fuentes heterogéneas o se van a querer combinar de maneras arbitrarias y no predefinidas.

3.2 Minería de datos

El *data mining*, es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto. Básicamente, el *data mining* surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales [9].

Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento; un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se ha recopilado o extraído los datos [6].

3.2.1 Definición de minería de datos

De las múltiples definiciones más o menos equivalentes que existen de *Data Mining* se pueden citar las siguientes:

- El Instituto SAS define el concepto de *Data Mining* como el proceso de Seleccionar (*Selecting*), Explorar (*Exploring*), Modificar (*Modifying*), Modelizar (*Modeling*) y Valorar (*Assessment*) grandes cantidades de datos con el objetivo de descubrir patrones

desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. Este proceso es resumido con las siglas SEMMA. El proceso de *Data Mining* es por tanto aplicable a lo largo de una amplia variedad de industrias y proporciona distintas metodologías de análisis según el tipo de problema que queramos analizar [10].

- Se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semiautomático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Por lo tanto, dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos entre otros problemas). Y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil [9].

Se puede decir que "en *data mining* cada caso es un caso". Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales:

1. **Determinación de los objetivos.** Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en *data mining*.

2. **Preprocesamiento de los datos.** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de *data mining*.

3. **Determinación del modelo.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

4. **Análisis de los resultados.** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente

determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

3.2.2 Aplicaciones

La integración de las técnicas de minería de datos en las actividades del día a día se está convirtiendo en algo habitual. Siendo un poco más concretos, a continuación incluimos una lista de ejemplos en algunas de las aéreas en las que se puede usar la minería de datos.

- Aplicaciones financieras y banca:
  - Obtención de patrones de uso fraudulento de tarjetas de crédito.
  - Determinación del gasto en tarjeta de crédito por grupos.
  - Cálculo de correlaciones entre indicadores financieros.
  - Identificación de reglas de mercado de valores a partir de históricos.
  - Análisis de riesgos en créditos.

- Análisis de mercado, distribución y, en general, comercio:
  - Análisis de la cesta de la compra (compras conjuntas, secuenciales, ventas cruzadas, señuelos, etc.).
  - Evaluación de campañas publicitarias.
  - Análisis de la fidelidad de los clientes. Reducción de fuga.
  - Segmentación de clientes.
  - Estimación de *stocks*, de costos, de ventas, etc.

- Seguros y salud privada:
  - Determinación de los clientes que podrían ser potencialmente caros.
  - Análisis de procedimientos médicos solicitados conjuntamente.
  - Predicción de que clientes contratan nuevas pólizas.
  - Identificación de patrones de comportamiento para clientes con riesgo.
  - Identificación de comportamiento fraudulento.
  - Predicción de los clientes que podrían ampliar su póliza para incluir procedimientos extras (dentales, ópticos...).

- Educación:
  - Selección o captación de estudiantes.
  - Detección de abandonos y de fracaso.
  - Estimación del tiempo de estancia en la institución.

- Procesos industriales:
  - Extracción de modelos sobre comportamiento de compuestos.
  - Detección de piezas con trabas. Modelos de calidad.
  - Predicción de fallos y accidentes.
  - Estimación de composiciones optimas en mezclas.
  - Extracción de modelos de costo.
  - Extracción de modelos de producción.

- Medicina:
  - Identificación de patologías. Diagnóstico de enfermedades.
  - Detección de pacientes con riesgo de sufrir una patología concreta.
  - Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
  - Recomendación priorizada de fármacos para una misma patología.

- Biología, bioingeniería y otras ciencias:
  - Análisis de secuencias de genes.
  - Análisis de secuencias de proteínas.
  - Predecir si un compuesto químico causa cáncer.
  - Clasificación de cuerpos celestes.
  - Predicción de recorrido y distribución de inundaciones.
  - Modelos de calidad de aguas, indicadores ecológicos.

- Telecomunicaciones:
  - Establecimiento de patrones de llamadas.
  - Modelos de carga en redes.
  - Detección de fraude.

- Otras áreas
    - Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo *spam*, gestión de avisos, análisis del empleo del tiempo.
    - Recursos Humanos: selección de empleados.
    - Web: análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los *logs* de un servidor web.
    - Turismo: determinar las características socioeconómicas de los turistas en un determinado destino o paquete turístico, identificar patrones de reservas, etc.
    - Tráfico; modelos de trafico a partir de fuentes diversas: cámaras, GPS...
    - Hacienda: detección de evasión fiscal.
    - Policiales: identificación de posibles terroristas en un aeropuerto.
    - Deportes: estudio de la influencia de jugadores y de cambios. Planificación de eventos.
    - Política: diseño de campañas políticas, estudios de tendencias de grupos, etc.
    - Desastres naturales: predicción y prevención de incendios forestales.

Todos estos ejemplos muestran la gran variedad de aplicaciones donde el uso de la minería de datos puede ayudar a entender mejor el entorno donde se desenvuelve la organización y, en definitiva, mejorar la torna de decisiones en dicho entorno.

3.2.3 Sistemas y herramientas de minería de datos

La diversidad de disciplinas que contribuyen a la minería de datos está dando lugar a una gran variedad de sistemas de minería de datos. Cada uno de ellos posee unas características apropiadas para realizar determinadas tareas o para analizar cierto tipo de datos. En esta sección presentamos varias clasificaciones de los sistemas y herramientas atendiendo a varios criterios (el modelo de datos que generan, el tipo de datos que minan, el tipo de técnica o el tipo de aplicación al que se pueden aplicar):

- **Tipo de base de datos minada:** teniendo en cuenta los diferentes modelos de datos podemos hablar de sistemas de minería de datos relacionales, multidimensionales, orientados a objetos, etc. Asimismo, atendiendo al tipo de datos manejados, hablamos de sistemas textuales, multimedia, espaciales o web.

- **Tipo de conocimiento minado:** también pueden tenerse en cuenta los niveles de abstracción del conocimiento minado: conocimiento generalizado (alto nivel de abstracción), a nivel primitivo (a nivel de filas de datos), o conocimiento a múltiples niveles (de abstracción). Por último, podemos igualmente distinguir entre los sistemas que buscan regularidades en los datos (patrones) frente a los que analizan las irregularidades (excepciones).

- **Tipo de funcionalidad y de técnica:** los sistemas de minería de datos se pueden clasificar basándose en su funcionalidad (clasificación, agrupamiento, etc.) o por los métodos de análisis de los datos empleados (técnicas estadísticas, redes neuronales, etc.).

- **Tipo de aplicación:** podemos distinguir dos clases de sistemas según la aplicación para la que se usan: los sistemas de propósito general y los sistemas específicos.

3.2.4 Tipos de modelos

La minería de datos tiene como objetivo analizar los datos para extraer conocimiento. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos.

En la práctica, los modelos pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que nos referiremos como variables independientes o predictivas.

Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos [6].

3.2.5 La minería de datos y el proceso de descubrimiento de conocimiento en bases de datos.

Todo este explosivo crecimiento de datos generó, a finales de los 80, la aparición de un nuevo campo de investigación que se denomino KDD (*Knowledge Discovery in Databases*). Bajo estas siglas se esconde, tal y como sugiere *Fayyad* et al. (1996), "el proceso no trivial de descubrimiento de patrones validos, nuevos, potencialmente útiles y comprensibles en grandes volúmenes de datos". El proceso de KDD ha servido para unir investigadores de aéreas en principio dispersas como Inteligencia Artificial, Estadística, Técnicas de Visualización, Matemáticas, Aprendizaje Automático o Bases de Datos en la búsqueda de técnicas eficientes y eficaces que ayuden a encontrar el potencial conocimiento que se encuentra inmerso en los grandes volúmenes de datos almacenados por las organizaciones diariamente [10].

Existen términos que se utilizan frecuentemente como sinónimos de la minería de datos, en el que puede destacar, la extracción o "descubrimiento de conocimiento en base de datos" (*Knowledge Discovery in Databases*, KDD). De hecho, en muchas ocasiones ambos términos se han utilizado indistintamente, aunque existen claras diferencias "entre los dos. Así, últimamente se ha usado el término KDD para referirse a un proceso que consta de una serie de fases, mientras que la minería de datos es solo una de estas fases.

Se define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". En esta definición se resumen cuales deben ser las propiedades deseables de conocimiento extraído:

- **Válido:** hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no solo para aquellos que han sido usados en su obtención.
- **Novedoso:** que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- **Potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- **Comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una

información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

Como se deduce de la definición anterior, el KDD es un proceso complejo que incluye no solo la obtención de los modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos, tal y como se refleja en la Figura 3.8.

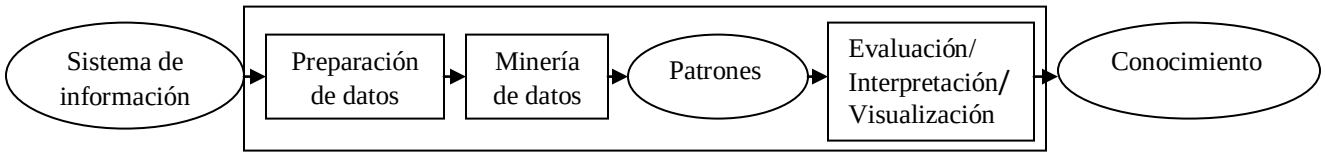

Figura 3. 8 *Proceso KDD* [6].

Así, los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso. Esta definición del proceso clarifica la relación entre el KDD y la minería de datos: el KDD es el proceso global de descubrir conocimiento útil desde bases datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos, al ser la fase de generación de modelos [6].

El objetivo de diseñar un modelo de proceso KDD es para llegar a un conjunto de pasos de procesamiento que pueden ser seguidos por los profesionales, cuando ejecutan sus proyectos. Este proceso debería ayudar a planificar el trabajo, y reducir los costos, detallando los procedimientos que se realizan en cada uno de los pasos. El modelo de proceso KDD proporciona una descripción completa de todos los pasos, desde la especificación del problema a la implementación de los resultados, como se muestran a continuación:

1. **La comprensión del dominio del problema:** En esta etapa se trabaja en estrecha colaboración con expertos en la materia para definir el problema y determinar los objetivos del proyecto, identificar a las personas clave, y aprender sobre las actuales soluciones al

problema. Se trata de aprender la terminología específica del dominio. Una descripción del problema, incluyendo sus restricciones y metas a ser alcanzadas.

2. **Comprensión de los datos:** Este paso incluye la recolección de datos de la muestra y decidir qué datos son necesarios, incluyendo su formato y tamaño. Si el conocimiento de fondo existe, algunos atributos pueden ser clasificados como más importante. A continuación, se debe verificar la utilidad de los datos en relación con los objetivos KDD. Los datos deben ser verificados para la integridad, redundancia, valores faltantes, la plausibilidad del valor de los atributos y cuestiones similares.

3. **Preparación de los datos:** Este es el paso clave del cual depende el éxito del proceso de descubrimiento de conocimiento, por lo general consume aproximadamente la mitad del esfuerzo de todo el proyecto. En este paso, decidir qué datos serán utilizados como insumo para las herramientas de minería de datos en el Paso 4.

4. **La minería de datos:** Este es otro paso clave, en el proceso de descubrimiento de conocimiento. A pesar de que la minería de datos es una herramienta que descubre nueva información, su aplicación conlleva menos tiempo que la preparación de datos. Este paso implica el uso de herramientas de minería de datos, y la selección de otras si es necesario. Las herramientas de minería de datos incluyen muchos tipos de algoritmos, como por ejemplo: los conjuntos, métodos Bayesianos, computación evolutiva, aprendizaje automático, redes neuronales, agrupación, y las técnicas de preprocesamiento.
   Una de las mayores dificultades en este paso es que muchas herramientas de uso común no pueden ampliarse para ser aplicadas a un gran volumen de datos. Las herramientas están caracterizadas por un aumento lineal de su tiempo de ejecución, con el aumento de datos dentro de una cantidad fija de memoria disponible. La mayor parte de las herramientas de DM no son escalables, pero hay ejemplos de herramientas que si lo son entre las que podemos encontrar: agrupamiento, aprendizaje automático, de reglas de asociación.

5. **Evaluación de los conocimientos descubiertos:** Este paso incluye la interpretación de los resultados por los expertos, evaluando la información si realmente es novedosa e interesante, y se lleva un control de impacto del conocimiento descubierto. Sólo los modelos aprobados (resultados de la aplicación de muchas herramientas de minería de datos y los métodos de pre-procesamiento) se mantienen. La todo el proceso de KDD puede

volver a examinar para identificar las acciones alternativas que se podrían tomar para mejorar los resultados.

6. **Utilizando el conocimiento descubierto:** Este paso está enteramente en manos de los propietarios de la base de datos. Se compone de la planificación de dónde y cómo el conocimiento descubierto se utilizará. El área de aplicación en el dominio actual debería extenderse a otros ámbitos dentro de una organización. Se debe crear un plan para monitorear el conocimiento descubierto y documentar el proyecto.

La característica importante del proceso de KDD es el tiempo relativo dedicado a completar cada uno de los pasos. Se estima que alrededor del 20% del esfuerzo se dedica a la determinación de objetivos de negocio, alrededor del 60% en la preparación de datos, y un 10% para la minería de datos y el análisis del conocimiento y la asimilación de conocimientos pasos, respectivamente. Otros autores muestran que alrededor del 15 al 25% del tiempo del proyecto se dedica a la etapa de DM. Por lo general se asume que aproximadamente el 50% del esfuerzo del proyecto está dedicado a la preparación de datos [12].

3.2.6 Tareas y métodos de minería de datos.

Una de las primeras cosas que debemos clarificar definitivamente antes de continuar es diferenciar una tarea de un método, así como destacar las tareas y métodos más relevantes. Una (un tipo de) tarea de minería de datos es un (tipo de) problema de minería de datos. Por ejemplo, "clasificar las piezas del proveedor Minatronix en óptimas, defectuosas reparables y defectuosas irreparables" es una tarea. Concretamente, el tipo de la tarea es clasificación. Esta tarea, por ejemplo, se podría resolver mediante arboles de decisión o redes neuronales, entre otros métodos. Estos son métodos o técnicas que permiten resolver tareas. Es muy importante distinguir el problema de los métodos para solucionarlo. Pasemos a ver, en primer lugar, las tareas y, posteriormente, los métodos más importantes.

En primer lugar, para definir las tareas debemos definir el conjunto de ejemplos con los que se van a tratar. Definamos E como el conjunto de todos los posibles elementos de entrada. Las instancias posibles dentro de E generalmente se representan como un conjunto de valores para una serie de atributos (sean nominales o numéricos). Es decir $E = A_1 \times A_2 \times ... \times A_n$ y un ejemplo e es una

tulpa <a_1, $a_{2,\ldots,}$ a_n> tal que a_i, ϵ A_i. Como veremos en próximos apartados, esta presentación tabular es la más habitual (pares atributos-valor), pero no la única. Veamos a continuación las tareas más importantes en minería de datos.

- **Predictivas:** se trata de problemas y tareas en los que hay que predecir un o más valores para uno o más ejemplos. Los ejemplos en la evidencia van acompañados de una salida (clase, categoría o valor numérico) o un orden entre ellos. Dependiendo de cómo sea la correspondencia entre los ejemplos y los valores de salida y la presentación de los ejemplos podemos definir varias tareas predictivas:

  - **Clasificación (o discriminación):** los ejemplos se presentan como un conjunto de pares de elementos de dos conjuntos, δ= { <*e, s*> : *e* ϵ *E*, s ϵ S }, donde *S* es el conjunto de valores de salida. Los ejemplos *e,* al ir acompañados de un valor de S, se denominan comúnmente ejemplos etiquetados <*e, s*> y, en consecuencia, δ se denomina conjunto de datos etiquetado. El objetivo es aprender una función
    λ*: E*→*S,* denominada clasificador, que represente la correspondencia existente en los ejemplos, es decir, para cada valor de *E* tenemos un único valor para *S.* Además, S es nominal, es decir, puede tomar un conjunto de valores c_1, c_2,..., c_m, denominados clases. La función aprendida será capaz de determinar la clase para cada nuevo ejemplo sin etiquetar, es decir dará un valor de S para cada valor de *e.*

- **Descriptivas:** los ejemplos se presentan como un conjunto δ= { *e : e* ϵ *E* }, sin etiquetar ni ordenar de ninguna manera. El objetivo, por tanto, no es predecir nuevos datos sino describir los existentes. No obstante, vamos a ver las tareas descriptivas mas delimitadas:

  - **Agrupamiento *(clustering):*** el objetivo de esta tarea es obtener grupos o conjuntos entre los elementos de δ, de tal manera que los elementos asignados al mismo grupo sean similares. Lo importante del agrupamiento respecto a la clasificación es que son precisamente los grupos y la pertenencia a los grupos lo que se quiere determinar y, a priori, no se sabe ni como son los grupos ni cuantos hay. En algunos casos se puede proporcionar el número de grupos que se desea obtener. Otras veces, este número se determina por el algoritmo de agrupamiento, según las características de los datos. La función a obtener es idéntica a la de la clasificación, λ*: E* → *S,* con la diferencia de que los valores de S y sus miembros se crean o inventan, durante el proceso de aprendizaje.

El agrupamiento se conoce muy frecuentemente por su término en ingles: *clustering*. También se pueden definir variantes para realizar un agrupamiento suave u obtener estimadores de probabilidad de agrupamiento, que proporcionan más flexibilidad y posibilidades a la hora de interpretar y trabajar con los grupos formados, o permiten construir taxonomías o agrupamientos jerárquicos.

- **Reglas de asociación:** Dados los ejemplos del conjunto $E = A_1 \times A_2 \times ... \times A_n$ una regla de asociación se define generalmente de la siguiente forma: "si $A_i = a \wedge A_j = b \wedge .... \wedge A_k = h$ entonces $A_r = u \wedge A_s = v \wedge .... \wedge A_z = w''$, donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributo. La regla anterior está orientada, es decir es una regla de asociación direccional. En realidad se buscan generalmente conjuntos de reglas de asociación, es decir más de una regla de asociación. A veces, conseguimos un conjunto tan bueno de reglas de asociación que si nos centramos en un solo atributo en la parte derecha de reglas orientadas podemos llegar a tener casi un clasificador.

Cada una de las tareas anteriores, como cualquier problema, requiere métodos, técnicas o algoritmos para resolverlas. Una de las cosas que sorprenden a los recién llegados a la minería de datos es que, además de que, lógicamente, una tarea puede tener muchos métodos diferentes para resolverla, tenemos que el mismo método (o al menos el mismo tipo de técnica) puede resolver un gran número de tareas.

Veamos brevemente los tipos de técnicas existentes para llevar a cabo las tareas anteriores. La relación que se muestra a continuación solo pretende dar una reseña de la variedad de técnicas existentes.

- **Técnicas algebraicas y estadísticas:** se basan, generalmente, en expresar modelos y patrones mediante formulas algebraicas, funciones lineales, funciones no lineales, distribuciones o valores agregados estadísticos tales como medias, varianzas, correlaciones, etc. Frecuentemente, estas técnicas, cuando obtienen un patrón, lo hacen a partir de un modelo ya predeterminado del cual, se estiman unos coeficientes o parámetros. Algunos de los algoritmos más conocidos dentro de este grupo de técnicas son la regresión lineal (global o local), la regresión logarítmica y la regresión logística.

Los discriminantes lineales y no lineales, basados en funciones predefinidas, es decir discriminantes paramétricos, entran dentro de esta categoría.

- **Técnicas bayesianas:** se basan en estimar la probabilidad de pertenencia (a una clase o grupo), mediante la estimación de las probabilidades condicionales inversas o a priori, utilizando el teorema de Bayes. Algunos algoritmos muy populares son el clasificador bayesiano Naive, los métodos basados en máxima verisimilitud y el algoritmo EM. Las redes bayesianas generalizan las topologías de las interacciones probabilísticas entre variables y permiten representar gráficamente dichas interacciones.

- **Técnicas basadas en conteos de frecuencias y tablas de contingencia:** estas técnicas se basan en contar la frecuencia en la que dos o más sucesos se presenten conjuntamente. Cuando el conjunto de sucesos posibles es muy grande, existen algoritmos que van comenzando por pares de sucesos e incrementando los conjuntos solo en aquellos casos que las frecuencias conjuntas superen un cierto umbral. Ejemplos de estos algoritmos son el algoritmo "Apriori" y similares.

- **Técnicas basadas en arboles de decisión y sistemas de aprendizaje de reglas:** son técnicas que, además de su representación en forma de reglas, se basan en dos tipos de algoritmos: los algoritmos denominados "divide y vencerás", como el 1D3/C4.5 o el CART, y los algoritmos denominados "separa y vencerás", como el CN2.

- **Técnicas relacionales, declarativas y estructurales:** la característica principal de este conjunto de técnicas es que representan los modelos mediante lenguajes declarativos, como los lenguajes lógicos, funcionales o lógico-funcionales. Las técnicas de ILP (programación lógica inductiva) son las más representativas y las que han dado nombre a un conjunto de técnicas denominadas minería de datos relacional.

- **Técnicas basadas en redes neuronales artificiales:** se trata de técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido. Existen innumerables variantes de organización: perceptrón simple, redes multicapa, redes de base radial, redes de Kohonen, etc., el más conocido es el de retro-propagación (backpropagation).

- **Técnicas basadas en núcleo y maquinas de soporte vectorial:** se trata de técnicas que intentan maximizar el margen entre los grupos o las clases formadas. Para ello se basan en unas transformaciones que pueden aumentar la dimensionalidad. Estas transformaciones se llaman núcleos (*kernels*). Existen muchísimas variantes, dependiendo del núcleo utilizado y de la manera de trabajar con el margen.

- **Técnicas estocásticas y difusas:** bajo este paraguas se incluyen la mayoría de las técnicas que, junto a las redes neuronales, forman lo que se denomina computación flexible (*soft computing*). Son técnicas en las que o bien los componentes aleatorios son fundamentales, como el *simulated annealing*, los métodos evolutivos y genéticos, o bien al utilizar funciones de pertenencia difusas (*fuzzy*).

- **Técnicas basadas en casos, en densidad o distancia:** son métodos que se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos (los casos más similares), de una manera más sofisticada, mediante la estimación de funciones de densidad. Además de los vecinos más próximos, algunos algoritmos muy conocidos son los jerárquicos, como *Two-step* o COBWEB, y los no jerárquicos, como K medias.

Además de todo lo anterior existen multitud de híbridos que dificultan más aún realizar una taxonomía razonable [6].

3.3 Software para *Data Mining*

Se pueden encontrar tanto en el ámbito comercial como en el académico una serie de entornos software diseñado para dar soporte al ejercicio de la minería de datos. La mayoría de estos sistemas de software integran en un mismo entorno capacidades para el procesado de datos, diferentes modelos de análisis, facilidades para el diseño de experimentos y soporte gráfico para la visualización de resultados. Su manejabilidad no se halla condicionada a que el usuario posea conocimientos de programación, ya que existe una interfaz que facilita la interacción entre el usurario y la herramienta.

Entre los sistemas más utilizados se pueden nombrar: SPSS, *Clementine*, WEKA, *Kepler*, DMS, *DBMiner*, YALE, *DB2 Intelligent Miner*, *SAS Enterprise Miner*, *STATISTICA* y *Data Miner*, entre otros.

3.3.1 WEKA (*Waikato Environment for Knowledge Analysis*)

Es un entorno para experimentación de análisis de datos, desarrollada por un equipo de investigadores de la universidad de Waikato (Nueva Zelanda), que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Para ello se requiere únicamente que los datos a analizar se almacenen en un formato conocido como ARFF (*Attribute Relation File Format*).

WEKA es un software de libre distribución, bajo licencia GNU, desarrollado en Java. Está construido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos.

WEKA se caracteriza por:

- **Preprocesado de datos:** selección de atributos, discretización, tratamiento de valores desconocidos y transformación de datos numéricos.

- **Modelos de aprendizaje:** árboles de decisión (J4.8, versión propia del algoritmo C4.5), tablas de decisión, vecinos más próximos, máquinas de vectores de soporte, reglas de asociación, métodos de agrupamiento y modelos combinados.

- **Visualización:** la interfaz gráfica se compone de diversos entornos. El entorno *Explorer* permite controlar todas las operaciones anteriores. El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. El entorno *Experimenter* facilita el diseño y la realización de experimentos complejos. El proceso global de minería de datos en

WEKA se acelera considerablemente gracias al entorno *KnowledgeFlow* que, de una forma gráfica y a modo de flujos de operaciones, permite definir la totalidad del proceso.

3.4 *Microsoft SQL Server*

Microsoft SQL Server es un sistema para la gestión de bases de datos producido por *Microsoft* basado en el modelo relacional. Sus lenguajes para consultas son T-SQL y ANSI SQL. **Microsoft SQL Server** constituye la alternativa de *Microsoft* a otros potentes sistemas gestores de bases de datos como son Oracle, PostgreSQL o MySQL.

Microsoft SQL Server se caracteriza por:

- Soporte de transacciones.
- Escalabilidad, estabilidad y seguridad.
- Soporta procedimientos almacenados.
- Incluye también un potente entorno gráfico de administración, que permite el uso de comandos DDL y DML gráficamente.
- Permite trabajar en modo cliente-servidor, donde la información y datos se alojan en el servidor y los terminales o clientes de la red sólo acceden a la información.
- Además permite administrar información de otros servidores de datos [13].

3.4.1 *Microsoft Integration Service*

Microsoft Integration Services. Es una plataforma integrada en *SQL Server 2008* para la creación de soluciones de integración de datos en base a procesos de extracción, transformación y carga, permitiendo controlar y configurar las extracciones y transformaciones de una manera versátil, de tal manera que se optimice los recursos de hardware y tiempos de ejecución reduciendo el impacto en los sistemas de origen de los datos.

Integration Services permite integrar datos entre aplicativos o en el procesamiento de información para *Datawarehouse* y herramientas de Inteligencia de Negocios. Los requerimientos

principales para aplicación de *Integration Services* en procesamiento de datos para *Datawarehouse* son los siguientes:

- La actualización periódica programada de las bases de datos.
- El envío de mensajes de correo electrónico como respuesta a eventos.
- El diseño de procesos ETL mediante una herramienta gráfica.
- El registro de LOGs de ejecución de procesos para administración y control.
- La depuración y mantenimiento de repositorios de datos.

Integration Services contiene herramientas para la creación y administración de paquetes. Se define como paquete a un conjunto de pasos o tareas de ejecución secuencial o en paralelo, los pasos son tareas específicas en el ciclo de extracción transformación y carga de datos. Los pasos de cada paquete pueden configurarse según un conjunto variado de tareas y transformaciones integradas, reduciendo la complejidad y el tiempo de programación al crear soluciones.

Integration Services es compatible con todos los proveedores estándar de acceso a datos como OLEDB, ODBC, acceso nativo de *SQL Server*, Microsoft Excel, acceso a archivos planos etc. normalmente en un *Datawarehouse* los diversos sistemas entregan información en diferentes formatos por lo que esta es una característica técnica importante en la implementación de un *Datawarehouse*.

3.5 Conclusiones

El marco teórico antes descrito, amplia los conocimientos acerca de DataWarehouse, minería de datos y de las herramientas que son utilizadas para el desarrollo de esta tesis. Con el marco teórico se tiene un panorama de los pasos a seguir para el desarrollo de la tesis.

4

Análisis y diseño

En el presente capítulo se describe la metodología a utilizar para el diseño, construcción e implementación del almacén de datos para el Sistema para la Toma de Decisiones de Incendios Forestales del Estado de Tlaxcala. También se detallan las técnicas de Minería de Datos que serán aplicadas a dicho Sistema.

4.1 Planteamiento y requerimientos

En base a los problema plateado en la sección 1.1.1 y considerando los objetivos que se desean alcanzar, mismos que fueron especificados en la sección 1.1.2, se determina que la metodología a seguir es el proceso KDD, el cual fue planteado en la sección 3.2.5. como se muestra a continuación:

1. **Integración y recopilación**

    - Integrar y recopilar los datos de Bases de Datos y otras fuentes muy diversas tanto internas como externas.

2. **Selección, limpieza y transformación**

   - En este proceso se debe eliminar el mayor número posible de datos erróneos o inconsistentes e irrelevantes.
   - La selección y la limpieza pueden acompañarse de la transformación de atributos, obteniendo como resultado un conjunto de filas llamado vista minable.
   - La vista minable integra datos de diferentes fuentes, los limpia, selecciona y transforma, con el fin de prepararlos para la modelización.

3. **Minería de Datos**

   - Una vez recolectados los datos de interés, se puede decidir qué tipo de patrón se quiere descubrir, donde el tipo de conocimiento que se desea extraer marcará claramente la técnica de minería de datos a utilizar.

4. **Evaluación e interpretación**

   - Se evalúan y analizan los patrones por expertos, de ser necesario se realizan las fases anteriores.

5. **Difusión**

   - Se hará uso del nuevo conocimiento y será distribuido a todos los posibles usuarios.

Para ilustrar las fases ya explicadas se muestra la figura 4.1.

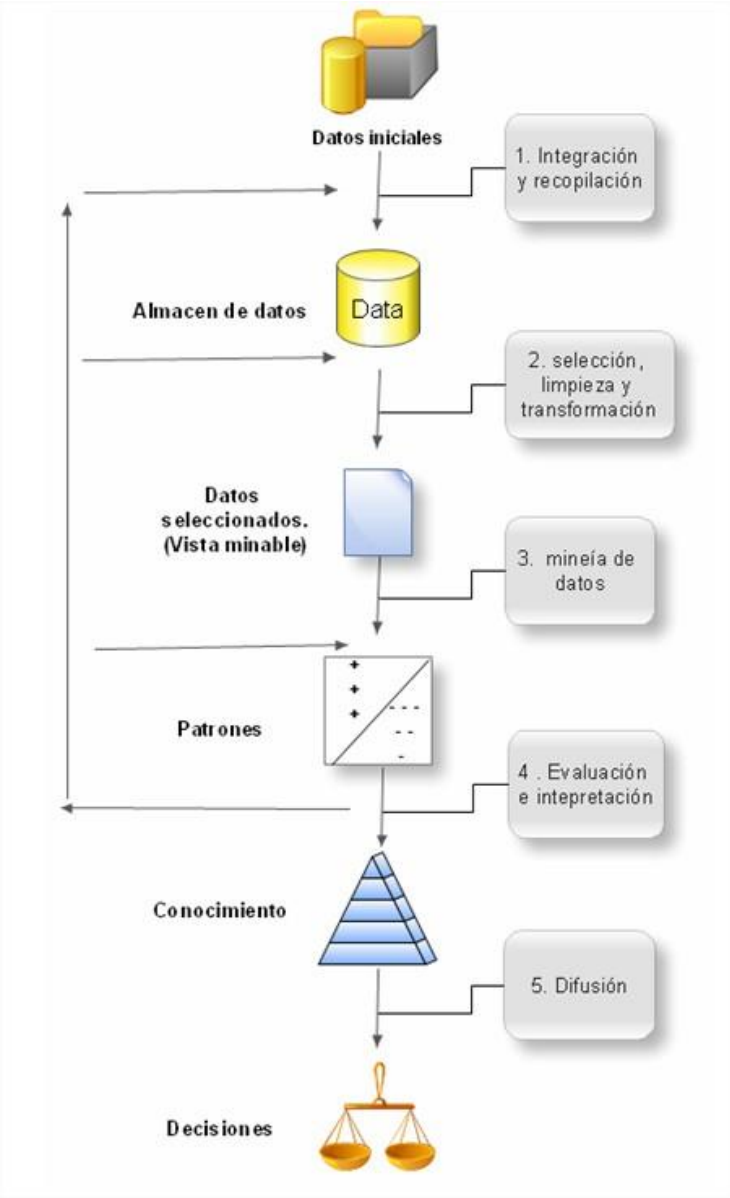

Figura 4. 1 *Fases del proceso de descubrimiento en bases de datos, KDD* [15].

4.2 Fase de integración y recopilación

En esta sección se describe la integración de los datos suministrados por CONAFOR. Los cuales han sido recabados del año 1999 a la fecha. Entre los datos recabados tenemos fecha incendio, coordenadas geográficas, municipio, predio, paraje, causa, horario de incendio, superficie afectada (arbolado renuevo, arbolado adulto, no arbolado matorrales y arbustos y no arbolado

pastizales) y personal participante (SEMARNAP, SEDENA, CGE, otras instituciones y voluntarios), los cuales se muestran en la tabla 4.1.

De los resultados obtenidos de la integración de datos, sobre los incendios forestales registrados en Tlaxcala se obtuvo una lista de las causas que los originan, la cual se aprecia en la tabla 4.2. De igual manera se obtiene la clasificación de la vegetación que se ve afectada, siendo dividida ésta en arbolado y no arbolado, ver tabla 4.3.

Tabla 4. 1 *Datos recolectados por CONAFOR*

Atributos
Fecha incendio
Coordenadas geográficas
Municipio
Predio
Paraje
Causa
Horario de incendio
Superficie afectada
Personal participante

Tabla 4. 2 *Causas registradas por CONAFOR*

Código	Causa
1	Actividades agropecuarias
2	Actividades forestales
3	Otras actividades productivas (IND, MAQ, ETC)
4	Limpia de derechos de vía
5	Fumadores
6	Fogatas de paseantes
7	Quema de basureros
8	Litigios
9	Rencillas
10	Para obtener autorizaciones de aprovechamientos forestales
11	Cazadores furtivos
12	Descargas eléctricas
13	Cultivos ilícitos
14	Ferrocarril
15	No determinadas

Tabla 4. 3 *Tipo de vegetación afectada*

Vegetación
Arbolado renuevo
Arbolado adulto
No arbolado matorrales y arbustos
No arbolado pastizales

Los datos suministrados por CONAFOR corresponden a 2389 registros, con sus respectivas mediciones.

4.3 Fase de selección, limpieza y transformación

La calidad del conocimiento obtenido mediante algoritmos de Minería de datos, no solo depende de este, sino también de la calidad de datos minados. Por dicha razón, se debe especificar la nomenclatura y transformación que deben tener los datos antes de almacenarlos en el *Data*

Warehouse. En esta etapa se verifican los datos que serán ingresados realizando una limpieza de estos en caso de ser necesario.

En este proceso se eliminaron 6 registros que no se encuentran dentro del rango de fechas proporcionado por CONAFOR. En los registros que no se tiene especificada una causa que origina el incendio, se asigna la causa 15 (No determinada). Para los campos de Municipio, Predio y Paraje los cuales son cadenas de caracteres se estandarizaron los datos eliminando acentos, espacios en blanco, signos de puntuación y poniéndolos en minúsculas. Este procedimiento se realizó mediante funciones de cadena de SQL Server 2008.

El en el campo horario de incendio se encuentra la hora de inicio y fin del incendio, dicho campo fue procesado con la finalidad de tener por separado la hora de inicio y fin del incendio esto para facilitar el tratamiento de los datos.

Una vez seleccionados los datos que serán almacenados en el *Data Warehouse*, se realiza la preparación de estos mediante la construcción automática de nuevos atributos, con ello facilitar el proceso de minería de datos. Por ejemplo los atributos de fecha fueron discretizados a tres rangos como se muestra en la figura 4.2.

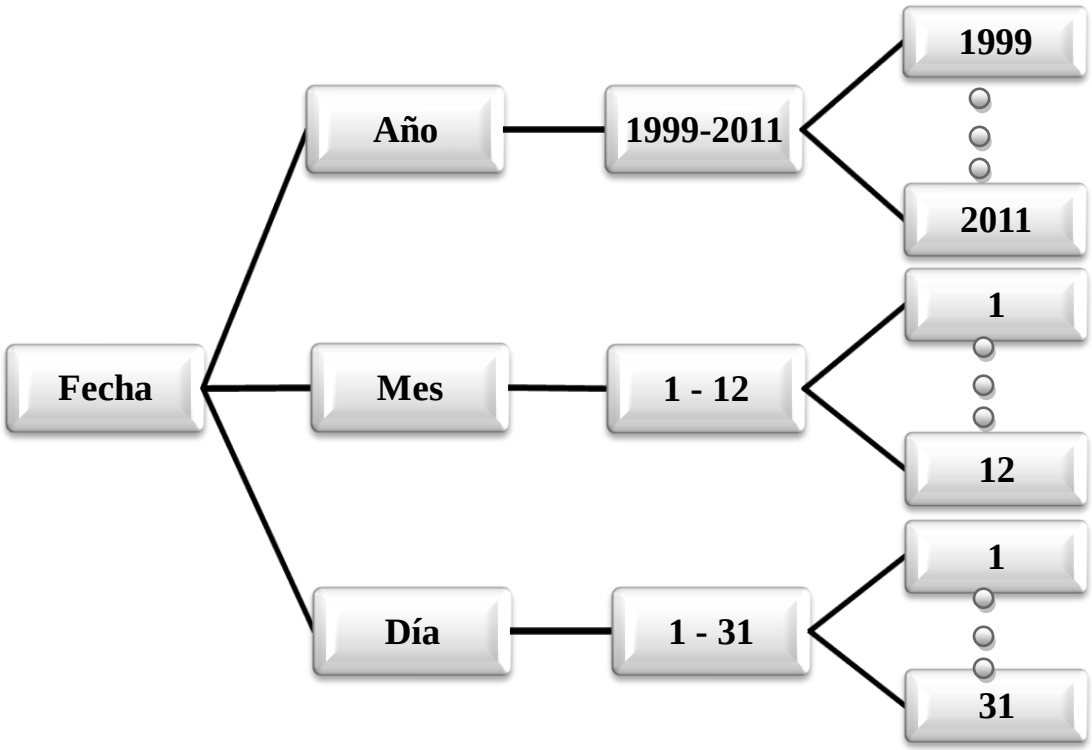

Figura 4. 2 *Ejemplo de discretización de fecha.*

4.4 Construcción del almacén de datos

Una vez realizado el proceso de selección, limpieza y transformación, se procede al cargado del *Data Warehouse* utilizando la herramienta de SQL *Server Business Intelligence Development Studio*. Para posteriormente realizar la construcción del cubo de datos, la cual es descrita a continuación.

Para diseñar una aplicación de *Business Intelligence* en SQL *Server*, primero debe crear un proyecto de *Analysis Services* en *Business Intelligence Development Studio*. En este proyecto, debe definir todos los elementos de la solución, empezando por una vista de origen de datos. Para crear una vista de origen de datos se deben realizar las siguientes tareas:

1. Agregar un proyecto de *Analysis Services* al proyecto de *Integration Services*, que fue utilizado para el proceso de selección, limpieza y transformación.
2. Definir un origen de datos que el proyecto utilizará, en el cual se define la información de cadena de conexión que se utilizará para establecer la conexión con el origen de datos.
3. Tras definir los orígenes de datos que utilizará en un proyecto de *Analysis Services*, el paso siguiente consiste en definir una vista del origen de datos para el proyecto. Dicha vista es una sola vista unificada de metadatos de las tablas y vistas especificadas que el origen de datos define en el proyecto. Almacenar metadatos en la vista de origen de datos permite trabajar con los metadatos durante el proceso de desarrollo sin ninguna conexión abierta con ningún origen de datos subyacente.

A partir de la base de datos, Incendios Forestales (la cual fue diseñada en SQL-Server 2008), se generó la vista de origen de datos, la cual se compone de las tablas: 1. Dim_Causa, 2. Dim_Tiempo, 3. Dim_Lugar y 4. Fact_Incendios (Ver figura 4.3).

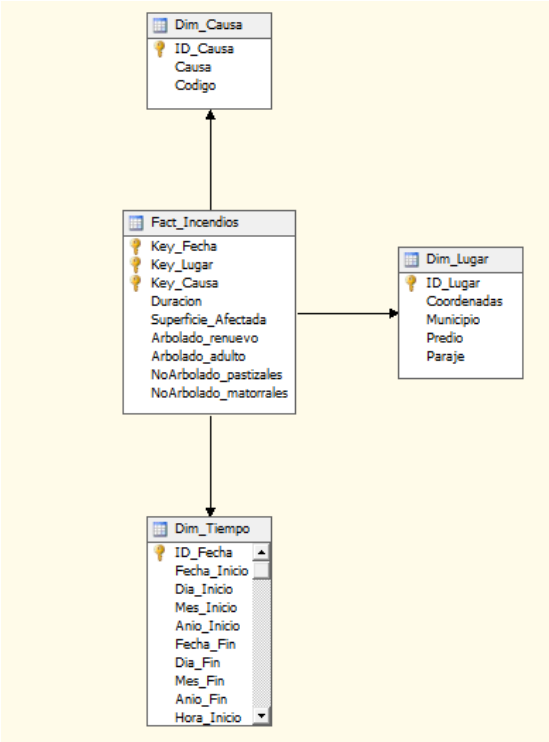

Figura 4. 3 *Vista de origen de datos generada a partir de la Base de Datos Incendios Forestales.*

Como se puede observar en la figura 4.3, la tabla de hechos o tabla principal es la que lleva por nombre Fact_Incendios, y es la encargada de recoger las características o atributos que van a permitir la relación con las dimensiones: Tiempo, Lugar y Causa.

Una vez que se ha definido una vista del origen de datos, se procede a definir e implementar el cubo, mediante la realización de las siguientes tareas:

1. Se definen las dimensiones, mediante el asistente para dimensiones. El resultado final se puede apreciar en la figura 4.4. la cual muestra las dimensiones que serán utilizadas en el cubo.
2. Para definir el cubo se usa el asistente para cubos, el cual ayudará a definir los grupos de medida y las dimensiones de un cubo. Una vez terminado el cubo se genera el diagrama que se observa en la figura 4.5.

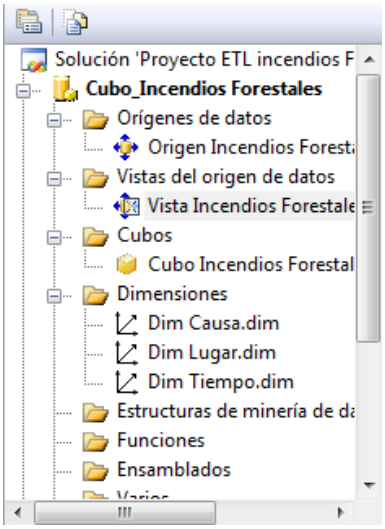

Figura 4. 4 *Dimensiones definidas.*

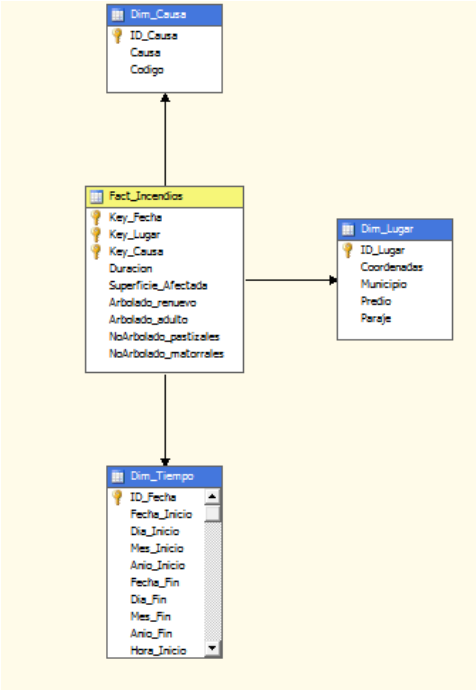

Figura 4. 5 *Tablas del almacén de datos*

4.5 Mantenimiento del almacén de datos

El mantenimiento o actualización del DW asegura contar con datos actualizados. Existen dos formas de refrescar los datos: la primera es llevar los datos al DW segundos después de que las fuentes fueron actualizadas. La segunda es acumulando y almacenando los datos ya integrados y transformados, en un sitio intermedio para de forma periódica pasar la información al DW. La actualización se puede realizar de manera incremental o recalculando todos los datos.

La actualización de un DW está considerado como un problema difícil debido a las siguientes razones: primero, el volumen de datos almacenado en el DW es muy grande y crece cada vez más. Segundo, la actualización debe ser accesible a los diferentes cambios de ejecución del DW. Finalmente, este proceso engloba transacciones que por lo regular acceden a múltiples datos, lo que implicaría contar con cálculos que pueden convertirse en complejos ya que producirían un alto nivel de agregación.

En este caso la actualización va a ser de forma periódica, mediante la utilización de *Agent SQL Server* el cual automatiza y programa la ejecución de SQL *Server Integration Services* mediante la ejecución de trabajos.

Para configurar el *Agent SQL Server* para automatizar la ejecución de un paquete del SQL *Server Integration Services* se deben seguir los siguientes pasos:

1. Abrir SQL *Server Management Studio* y conectar con el motor de base de datos de SQL *Server*. La cuenta que ejecuta el paquete como un paso de trabajo debe tener los mismos permisos que una cuenta que ejecuta el paquete directamente, para tener acceso a cualquier recurso externo al que debe tener el paquete.

2. Crear una credencial, mediante el cuadro de diálogo nueva credencial. Se asigna el nombre de la credencial, se especifica la cuenta utilizada en las conexiones salientes (normalmente será una cuenta de usuario de Windows) finalmente se asigna una contraseña (Si se ha especificado una cuenta de usuario de Windows en Identidad, ésta será la contraseña de Windows).

3. Crear una cuenta *proxy* que ejecuta el paquete como un paso de trabajo del *Agent* SQL *Server*. En esta ocasión se hará mediante un cuadro de diálogo Nueva cuenta proxy en SQL *Server Management Studio*. En la página General del cuadro de diálogo Nueva cuenta de proxy, especifique el nombre del proxy, el nombre de la credencial, la descripción del nuevo proxy y especificar el subsistema para el que el proxy este habilitado (SQL *Server Integration Services Package*).

4. Iniciar *Agent SQL Server*, clic derecho en *Agent SQL Server* e iniciar, ver figura 4.6.

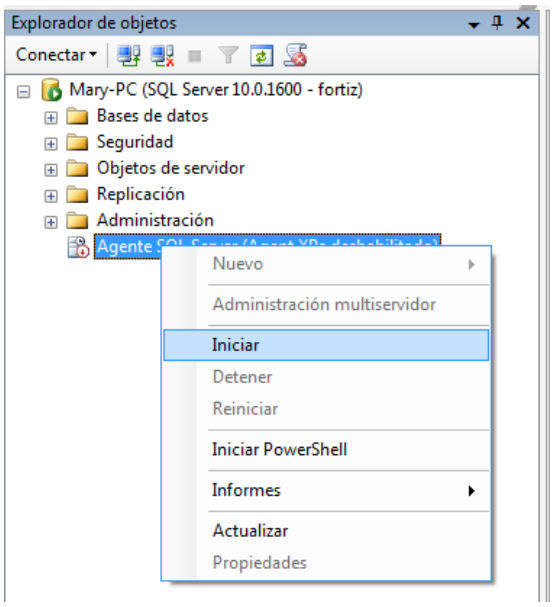

Figura 4. 6 *Ejecutar Agent SQL Server.*

5. Crear el trabajo. Para esto se expande *Agent SQL Server*, clic secundario en Trabajos y, a continuación, haga clic en Nuevo trabajo. Configurar los propiedades del trabajo en la página general del cuadro de diálogo (ver figura 4.7) especificar nombre de trabajo y confirmar los valores de propietario y categoría son exactos. Se selecciona la casilla habilitado para asegurar que el trabajo pueda ser programado.

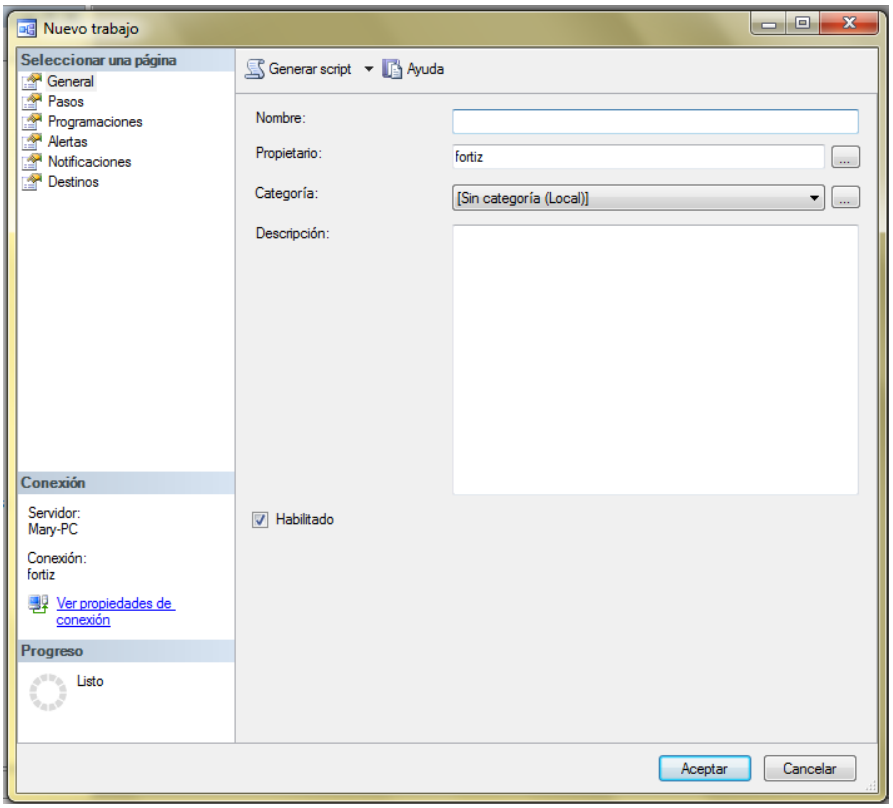

Figura 4. 7 *Cuadro de diálogo nuevo trabajo.*

6. Agregar un paso de trabajo en la página pasos, seleccionar nuevo paso, se da un nombre, selecciona el tipo de paquete (SQL *Server Integration Services Package*), en ejecutar como se selecciona la cuenta proxy antes creada, en la ficha General se selecciona el origen del paquete (ver figura 4.8) y se agrega un proveedor de registros para escribir entradas de registro para los eventos de una base de datos de SQL *Server*, usando un administrador de conexiones en el paquete (MARY-PC.IncendiosForestales.Fortiz conecta con la base de datos de SQL *Server*) ver figura 4.9.

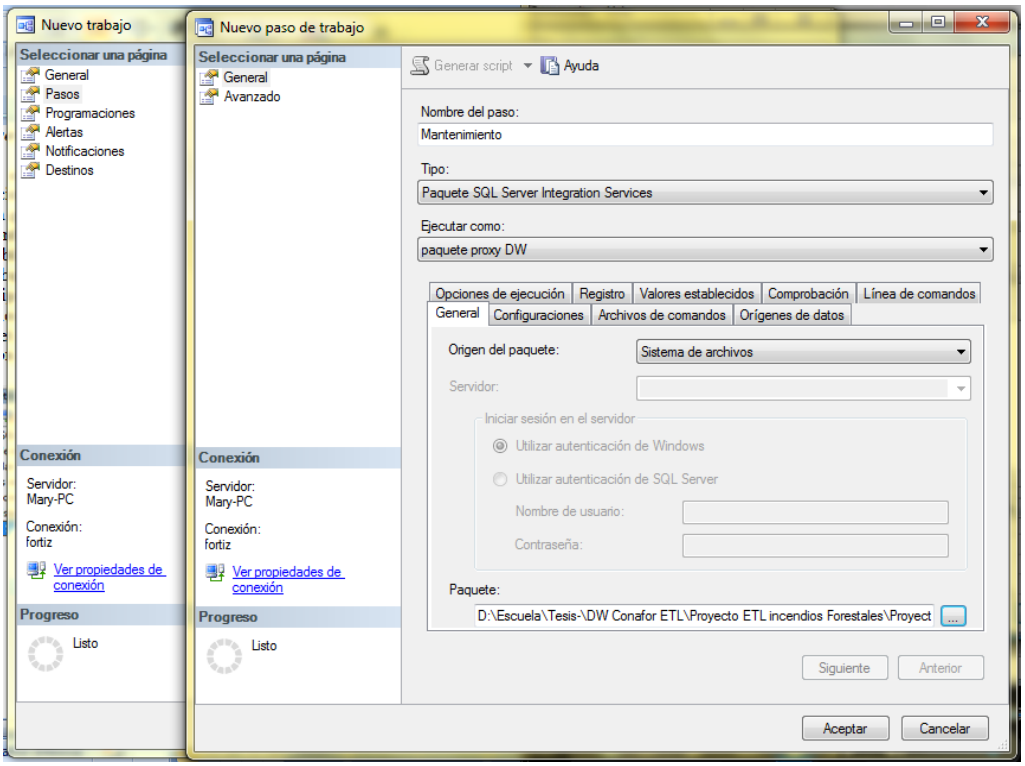

Figura 4. 8 *Cuadro de diálogo nuevo paso de trabajo*

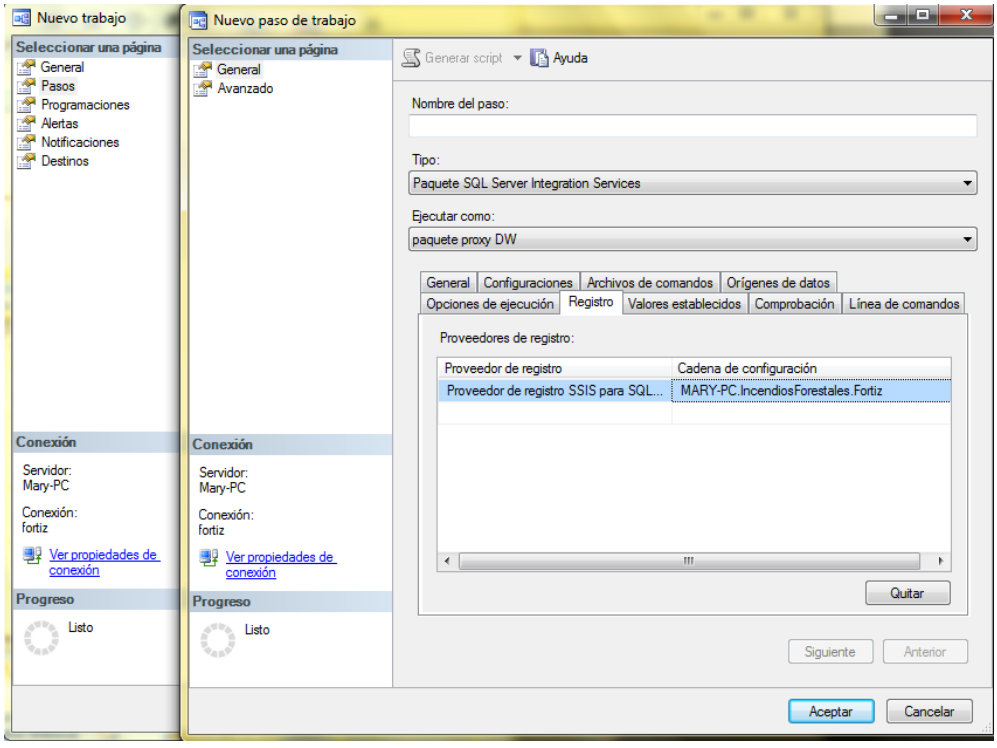

Figura 4. 9 *Agregar proveedor de registros.*

7. Crear una programación para el paso de trabajo. En la página de programaciones seleccionar nueva programación. En este cuadro de diálogo se especificara el nombre de la programación, y cuándo y con qué frecuencia se ejecutará. Para finalizar clic en aceptar ver figura 4.10.

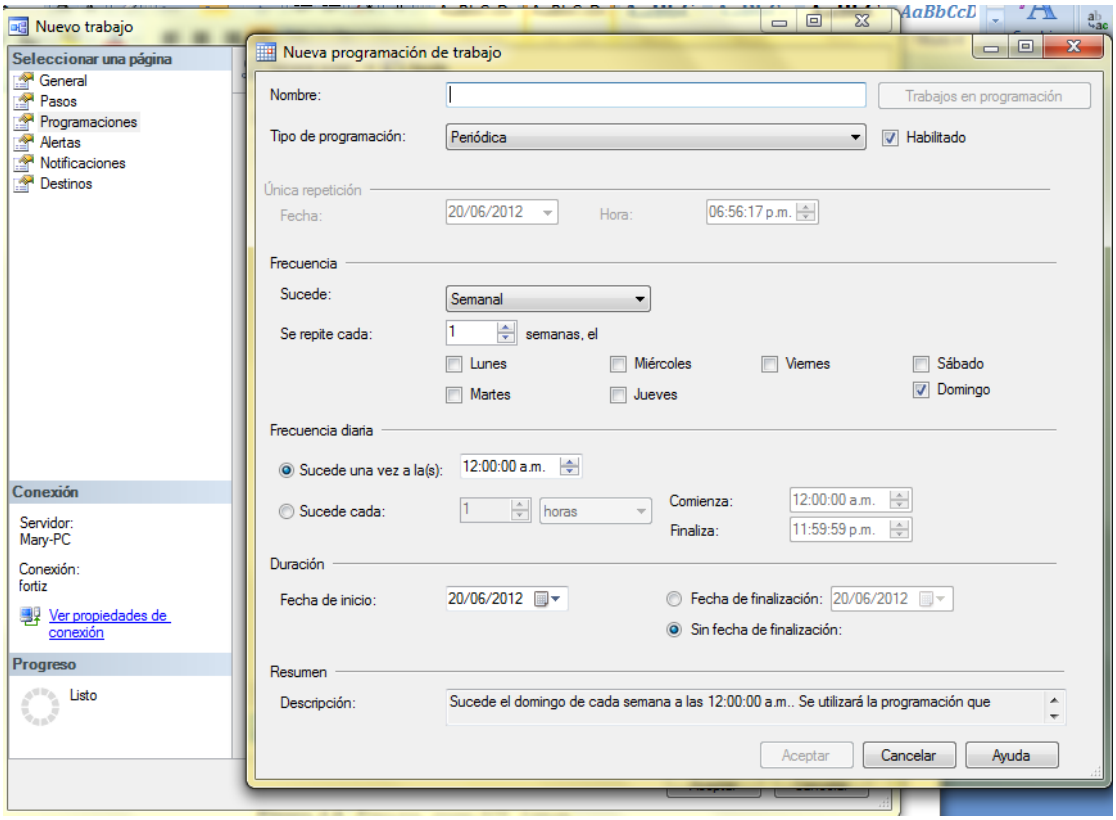

Figura 4. 10 *Programación de un paso de trabajo.*

El sistema ejecutará ahora automáticamente el paquete en una programación establecida. Usando la cuenta proxy que creamos pata ejecutar el trabajo, de igual manera se puede ejecutar el trabajo en cualquier momento fuera del marco de tiempo establecido por la programación.

4.6 Fase de minería de datos

Esta fase tiene como objetivo extraer y presentar conocimiento implícito, que pueda ser utilizado por el usuario.

Esto se realiza a partir de un modelo basado en un gran volumen de datos, que han sido recopilados y procesados para dicho fin. En estos datos se encuentran patrones, reglas y relaciones que no se aprecian a simple vista, y pueden utilizarse para explicar situaciones pasadas, entender los datos y hacer predicciones.

En los siguientes párrafos se describen las tareas utilizadas, para minar los datos que ya han sido procesados anteriormente:

4.6.1 Tareas elegidas de Minería de Datos

- **Agrupamiento (*clustering*):** Dicha tarea forma grupos tales que los objetivos de un mismo grupo son muy similares entre sí y, al mismo tiempo son muy diferentes a los objetivos de otros grupos**.** Por esta razón fue elegida para obtener los principales grupos de condiciones, bajo las cuales se produce un incendio, tales como mes, causa y municipio. En particular, para este caso se eligió el método ***SimpleKMeans.***

- **Clasificación:** Esta tarea es elegida con el objetivo de predecir la clase de las nuevas instancias para las que se desconoce la clase. Se eligió la técnica de arboles de decisión, en concreto el método **J48**, para conocer que municipios y en qué mes(es), se ven afectados por incendios forestales.

- **Asociación:** Con esta tarea se busca establecer relaciones asociativas entre los distintos datos de la Base de Datos. Las reglas de asociación no implican una relación causa-efecto, es decir, puede o no existir una causa para que los datos estén asociados. Por esta razón se eligió el método ***PredictiveApriori,*** para conocer la relación que existe entre los municipios, los meses y superficie afectada.

Las técnicas ya descritas, fueron aplicadas al modelo de datos, mediante el software WEKA.

4.7 Conclusiones

En este capítulo se presentó la metodología KDD la cual fue empleada, para el diseño, construcción e implementación del Almacén de Datos del sistema de Toma de Decisiones de Incendios Forestales. Se explicó a grandes rasgos la forma en que cada una de sus etapas modificó los datos.

De igual manera, se ofrece una explicación de las configuraciones, llevadas a cabo en SQL Server 2008, para el mantenimiento (actualización de datos) del Almacén de Datos.

Finalmente se describieron las tareas, técnicas y métodos de Minería de Datos que se emplearon para la obtención del conocimiento contenido en el Almacén de Datos antes construido.

5

Resultados

En el siguiente capítulo se muestran los resultados obtenidos, al ejecutar las técnicas de Minería de Datos, elegidas para la explotación del Almacén de Datos del Sistema para la toma de Decisiones de Incendios Forestales.

5.1 Agrupamiento (*clustering*)

Se aplicó el método de *SimpleKMeans* al conjunto de datos. En los resultados se observó que los factores que discriminan los grupos son los meses y los municipios. En estos municipios se propician incendios generalmente en los meses de febrero, marzo y abril, donde las condiciones climáticas son más propicias para la generación y propagación de incendios que el resto de meses. Otro aspecto de importancia que se observa, es el tipo de vegetación que se ve afectada, en su mayoría son pastizales seguidos de matorrales los cuales por sus características son fáciles a incendiarse y propagar el incendio con mayor rapidez.

Teniendo los municipios que se ven afectados por las causas ya descritas, se procede a analizar el comportamiento de sus parajes, para con ello identificar con mayor precisión las zonas de riesgo y bajo qué condiciones se generan. Una descripción de los resultados obtenidos al aplicar el método *SimpleKMeans* para los municipios se encuentran en la tabla 5.1. Las dimensiones de vegetación afectada se encuentran dadas en hectáreas.

Tabla 5. 1 *Resultados obtenidos al aplicar el método SimpleKMeans a los municipios.*

Código	**1**	**1**	**1**	**1**
Municipio	Teolocholco	Huamantla	Zitlaltepec	Chiautempan
Mes	4	2	4	3
Arbolado_Adulto	0.0051	0	3.8333	0.0296
Arbolado_Renuevo	0.0409	0.0398	2.3654	0.1091
NoArbolado_Matorrales	0.5678	0.5962	14.9038	0.9663
NoArbolado_Pastizales	1.4565	1.6573	16.266	2.6261

Como ya se mencionó se analizaran los predios de Teolocholco, Huamantla, Zitlaltepec y Chiautempan.

Para Teolocholco se aplica el método *SimpleKMeans* con 3 *clusters*, obteniendo como resultado que la causa es la "1", la cual corresponde a actividades agropecuarias y que afecta a los tres predios. Para San Luis los incendios se generan principalmente en los meses de Marzo y Abril, mientras que en Acxotla solo en e Marzo. En cuanto a la vegetación, se ven más afectado los pastizales, mientras el arbolado adulto tiene una afectación casi nula. La visualización de los datos ya descritos se observan en la tabla 5.2.

Tabla 5. 2 *Resultados obtenidos al aplicar el método SimpleKMeans al predio Teolocholco.*

Código	**1**	**1**	**1**
Municipio	San Luis	San Luis	Acxotla
Mes	3	4	3
Arbolado_Adulto	0	0	0.0909
Arbolado_Renuevo	0.0323	0.0366	0.0455
NoArbolado_Matorrales	0.6694	0.5244	0.3977
NoArbolado_Pastizales	1.5161	0.878	1.4943

Para el municipio de Huamantla, el predio con mayor afectación es Los Pilares con una amplia variedad de causas, entre las que tenemos causa 1, 2 y 6 correspondientes a actividades agropecuarias, actividades forestales y fogatas de paseantes respectivamente, observándose una mayor afectación en los meses de Marzo y Abril. La vegetación que se ve más afectada son los

pastizales, mientras que el arbolado adulto tiene una afectación nula. En la tabla 5.3 se visualiza los datos ya explicados.

Tabla 5. 3 *Resultados obtenidos al aplicar el método SimpleKMeans al predio Huamantla.*

Código	**2**	**6**	**1**
Municipio	Los Pilares	Los Pilares	Los Pilares
Mes	3	4	3
Arbolado_Adulto	0	0	0
Arbolado_Renuevo	0.1964	0.0656	0.5545
NoArbolado_Matorrales	0	0.225	0.3636
NoArbolado_Pastizales	2.3036	1.0906	2.9273

En el municipio de Zitlaltepec, se aplica el método *SimpleKMeans* con 2 *clusters*, en el resultado se observa que la causa 1 (actividades agropecuarias), afectando a los predios de Javier Mina y San Pablo en los meses de Abril y Marzo respectivamente. Como se puede observar en Javier Mina existe una afectación importante en pastizales y matorrales pero para el arbolado adulto la afectación es nula, para San Pablo la vegetación más afectada son pastizales. Para corroborar la información ya descrita se puede observar la tabla 5.4.

Tabla 5. 4 *Resultados obtenidos al aplicar el método SimpleKMeans al predio Juan Zitlaltepec.*

Código	**1**	**1**
Municipio	Javier Mina	San Pablo
Mes	4	3
Arbolado_Adulto	0	0
Arbolado_Renuevo	0.1667	0.0847
NoArbolado_Matorrales	1.5714	0.1949
NoArbolado_Pastizales	5.6176	2.0847

Para analizar los predios de Chiautempan se utilizó el método *SimpleKMeans* con 3 clusters. Se obtuvo como resultado que el predio de San Bartolo se ve afectado por la causa 1 (actividades agropecuarias) en el mes de Abril con una importante afectación en pastizales seguida de matorrales. Para el predio Tlalcuapan se ve afectado por la causa 1 (actividades agropecuarias) en Marzo, afectado principalmente los pastizales. Finalmente, en el predio Muñoztla los incendios

son generados por la causa 1 (actividades agropecuarias) en el mes de Marzo, afectando principalmente pastizales. Para tener una idea más clara sobre los resultados ya descritos analizar la tabla 5.5

Tabla 5. 5 *Resultados obtenidos al aplicar el método SimpleKMeans al predio Juan Chiautempan.*

Código	**1**	**1**	**1**
Municipio	San Bartolo	Tlalcuapan	Muñoztla
Mes	4	3	3
Arbolado_Adulto	0.0339	0	0
Arbolado_Renuevo	0.0678	0.0287	0.1173
NoArbolado_Matorrales	1.3305	0.2644	1.0816
NoArbolado_Pastizales	3.1102	1.1552	2.2194

Una vez analizados y explicados los resultados obtenidos con el método *SimpleKMeans*, se concluye que la principal causa de incendios en el estado de Tlaxcala es la uno, es decir, por actividades agropecuarias (causa 1), los cual indica que es necesario crear campañas de prevención de incendios orientadas a este sector. Por otro lado se tiene que la temporada alta de incendios se encuentra en el mes de marzo.

5.2 Clasificación

Los arboles de decisión fueron utilizados para visualizar la afectación que sufren los municipios, en determinado mes del año, teniendo como entradas los meses, municipios y superficie afectada, a este atributo se le aplico un filtro de discretización en cuatro intervalos.

El modelo resultante es robusto pero podemos obtener información muy valiosa.

En el árbol de la figura 5.1, se pueden apreciar los municipios que se ven afectados, en determinado mes y que superficie se ve afectada. Se observa que existen municipios que no tienen un gran índice de incendios (en base al análisis de agrupamiento antes explicado), pero su superficie forestal afectada es grande, entre dichos municipios podemos encontrar a Calpulalpan, Tlaxco, y Altzayanca, entre otros. En caso contrario se pudo observar que algunos de los municipios que han sido identificados con un alto índice de incendios, su superficie afectada es pequeña. Una

explicación a ese fenómeno podría ser su ubicación geográfica, vegetación, clima, materiales y combustibles, entre otros factores que influyen en la fácil propagación o extinción del incendio.

En base a estos resultados se propone al Gobierno del Estado de Tlaxcala, realizar campañas para la prevención de incendios, así como crear planes de contingencia para cada municipio.

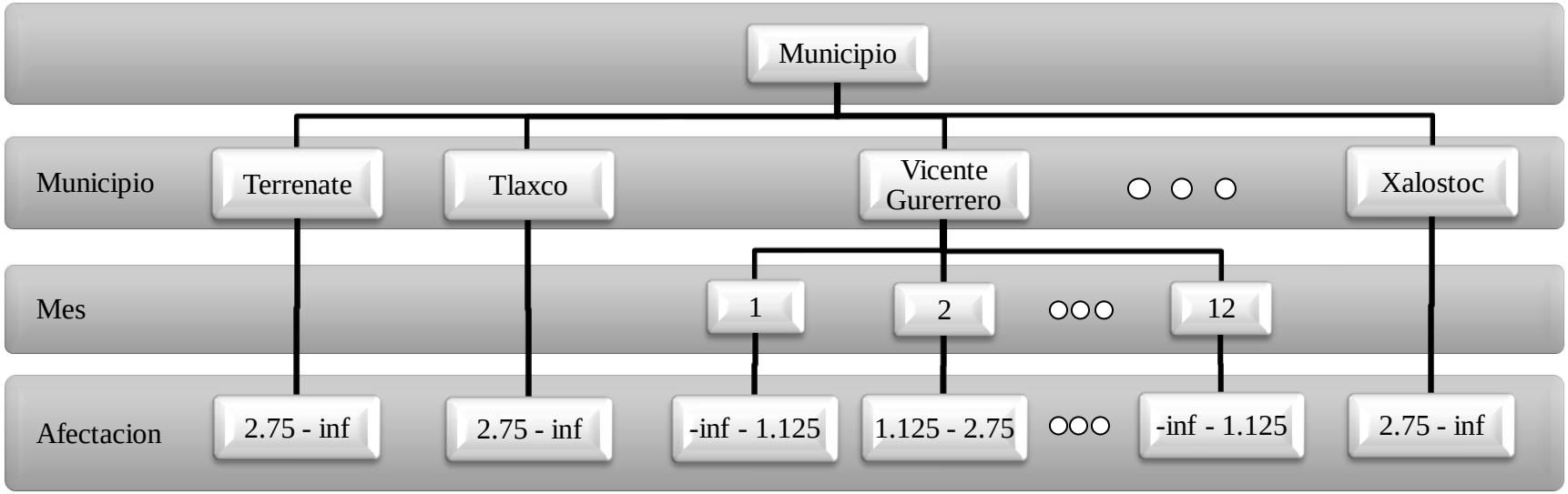

Figura 5. 1 *Muestra del árbol de decisión obtenido.*

5.3 Asociación

La técnica de asociación se empleó para analizar la relación existente entre el municipio, el mes y la superficie afectada. El algoritmo de asociación implementado en WEKA es el algoritmo ***PredictiveApriori,*** por simplicidad se aplicó un filtro de discretización al atributo superficie afectada en cuatro intervalos para explorar las relaciones más significativas. El algoritmo se ejecuta con los parámetros por defecto.

Para este caso, entre las reglas más significativas obtenidas se tienen:

1. Municipio=calpulalpan Mes_Inicio=5 9 ==> Superficie_Afectada='(2.75-inf)' 9 acc:(0.9873)
2. Municipio=xaltocan 7 ==> Superficie_Afectada='(2.75-inf)' 7 acc:(0.9793)
3. Municipio=altzayanca 17 ==> Superficie_Afectada='(2.75-inf)' 16 acc:(0.96005)
4. Municipio=terrenate Mes_Inicio=5 5 ==> Superficie_Afectada='(2.75-inf)' 5 acc:(0.95785)
5. Municipio=terrenate Mes_Inicio=3 4 ==> Superficie_Afectada='(2.75-inf)' 4 acc:(0.9337)
6. Municipio=juan cuamatzi Mes_Inicio=6 4 ==> Superficie_Afectada='(-inf-1.125]' 4 acc:(0.9337)
7. Municipio=panotla Mes_Inicio=5 4 ==> Superficie_Afectada='(2.75-inf)' 4 acc:(0.9337)
8. Municipio=xicohtencatl Mes_Inicio=1 4 ==> Superficie_Afectada='(2.75-inf)' 4 acc:(0.9337)
9. Municipio=huamantla Mes_Inicio=6 4 ==> Superficie_Afectada='(-inf-1.125]' 4 acc:(0.9337)

Estas reglas aportan información no tan trivial: el 98 % de los incendios generados en Calpulalpan en el mes de Mayo superan las 2.75 hectáreas afectadas, así el 97% de incendios generados en Xaltocan se ven afectadas más de 2.75 hectáreas.

Es significativo observar que los municipios catalogados con un alto índice de incendios, su área afectada es menor, tal es el caso de la regla 9, donde se dice que el 93% de los incendios en el Municipio de Huamantla en el mes de Junio muestran una afectación inferior a 1.125 hectáreas.

Con este análisis se concluye que hay municipios que no tienen una alta incidencia de incendios, pero con un grado de afectación alto, los cuales deben ser considerados para realizar campañas de prevención, de limpieza de suelos, capacitación de brigadas y creación de planes de contingencia.

5.4 Conclusiones

Los incendios forestales se consideran un problema importante tanto en el ámbito forestal como para la sociedad en general, por ello es evidente la necesidad de contar con información especialmente explícita sobre el fenómeno con el objeto de poder analizar posibles patrones de ocurrencia que contribuyan a optimizar las labores de vigilancia y protección de las zonas más afectadas. En este sentido, se desarrolla esta tesis con el apoyo de WEKA y métodos de análisis de datos que ayuden a la interpretación de la información.

En el presente proyecto se ha creado un Almacén de Datos, que contiene la información sobre los incendios forestales que se han registrado en el estado de Tlaxcala, para posteriormente ser analizados por herramientas de Minería de Datos. Esto ha permitido conocer a gran detalle el comportamiento de los incendios forestales en Tlaxcala. En términos generales se puedo concluir que:

- Se analizó el comportamiento de los incendios en cada mes del año. Con el objetivo de conocer cuál es o son los meses con mayor índice de incendios. Para ello se aplicó la tarea de agrupamiento, concluyendo que el mes con mayor índice de incendios es marzo seguido del mes de abril, donde las lluvias son escasas y se secan las tierras generando un escenario ideal para la generación y propagación de incendios.

- Se analizó el comportamiento de las causas que generan incendios, mediante la aplicación del método *SimpleKMeans*, concluyendo que los incendios generalmente son iniciados por actividades agropecuarias seguido de fogatas de paseantes, correspondientes a la causa 1 y la causa 6 respectivamente.

- Mediante un árbol de decisión generado por el método J48 se obtuvo una relación del municipio con la superficie afectada y en algunos casos esta relación se da por meses.

- Mediante reglas de asociación se identifican los municipios que tienen un alto grado de afectación en sus recursos forestales, así como en los que su afectación es menor, recalcando que estos municipios son los que tienen mayor incidencia de incendios.

6

Conclusiones y Trabajo a Futuro

Este capítulo presenta las conclusiones y comentarios finales del presente trabajo de investigación. Discute brevemente las lecciones aprendidas en durante su desarrollo, así como de las aportaciones alcanzadas. Por último ofrece un panorama del posible trabajo que pudiera seguir este proyecto en el futuro.

6.1 Conocimientos adquiridos en el desarrollo del proyecto

En el desarrollo de la tesis se han tenido que resolver distintos problemas, para ello ha sido necesario adquirir y recordar una serie de conocimientos en los siguientes temas:

- Excel debido a que la información proporcionada por CONAFOR se encuentra en este formato.
- SQL Server para la construcción del Almacén de Datos.
- WEKA para la aplicación de técnicas y por tanto la generación de modelos que reflejan distintos aspectos de los incendios forestales facilitando la comprensión del fenómeno.

Los conocimientos, a nivel de usuario informático en las herramientas referidas, adquiridos servirán para desempeñar múltiples tareas profesionales, entre ellas la ampliación del presente proyecto, el tratamiento de grandes volúmenes de información mediante SQL Server, la explotación

de información mediante WEKA, así como obtener conocimiento útil y novedoso de grandes volúmenes de información.

6.2 Aportaciones

Unas de las aportaciones que el presente trabajo de investigación ofrece, es la creación de modelos que establecen que la generación de incendios tiene una dependencia directa con el mes y causa. Es decir, la temporada alta de incendios es en el mes de marzo donde las lluvias son escasas, estos causados principalmente por actividades agropecuarias (causa 1), así como también, la superficie forestal, en promedio que se ve afectada por los incendios. Esta información es producto de la aplicación de técnicas de Minería de Datos.

Se ratificó la importancia y confiabilidad que tiene la información recolectada por CONAFOR.

La Minería de Datos permite descubrir patrones escondidos en vastas bases de datos. La combinación de esta tecnología con expertos en el tema, construye modelos que ayudan a crear planes estratégicos, para la prevención y sofocación de incendios.

6.3 Trabajo futuro

Entre los posibles caminos que pudiera seguir este trabajo de tesis en un futuro están el trabajar en la recolección de información tal como condiciones climáticas, profundidad de la materia orgánica, tipo de vegetación, elevación, pendientes, exposición, proximidad a caminos, cercanía de áreas agropecuarias, entre los más significativos. Esta nueva información facilitará la construcción de modelos predictivos que son necesarios para saber en qué zonas probablemente se generará un incendio dado sus condiciones físicas y climáticas, predecir un aproximado del área afectada y crear planes de contingencia adecuados para cada zona. Esto permitiría a los especialistas crear campañas de prevención y planes de contingencia adecuados para sofocar en menor tiempo un incendio y disminuir las hectáreas afectadas.

Otro posible plan a futuro, es la aplicación de otras técnicas de Minería de Datos. De esta manera enriquecer el conocimiento ya obtenido, manteniendo siempre el enfoque de la prevención, detección y sofocación de incendios.

Con los resultados obtenidos de este trabajo de tesis, se propone en particular, el monitoreo de los predios que se encuentran cerca de áreas agropecuarias y de zonas turísticas, para así tomar las medidas necesarias en la época alta de incendios.

6.4 Conclusiones finales

En esta tesis se llevó a cabo el análisis, diseño e implementación de un sistema para la toma de decisiones sobre Incendios Forestales.

La experiencia obtenida en el desarrollo del proyecto, permite concluir:

- El tener un conocimiento previo de la información permite identificar los problemas y necesidades de los incendios forestales.

- Se decide usar herramientas de software libre para los procesos de extracción y exploración por sus bajos costos.

- En base a las características propias de CONAFOR el uso de la metodología de Ralph Kimball resulta una solución eficaz en tiempo y recursos debido a que abarca la solución al problema en un corto plazo.

- Se diseña un modelo dimensional adecuado según la cantidad y profundidad de datos que posee el Almacén de Datos.

- El desarrollo de los procesos de extracción, transformación y carga son los apropiados según la información requerida.

- El uso de WEKA permite un manejo intuitivo y sencillo, sin requerir de un especialista informático.

Bibliografía

1. Sistema inteligente para el diseño asistido de planes de operaciones para la extinción de incendios forestales. SIADEX [en línea]: 2005-2011. [fecha de consulta: Julio 2011]. Disponible en:
   < http://noticias.universia.es/ciencia-nn-tt/noticia/2005/07/13/606808/apagar-fuego.html >

2. Sistema inteligente para el diseño asistido de planes de operaciones para la extinción de incendios forestales. SIADEX [Actualización]: 8 de junio, 2010. [fecha de consulta: Julio 2011]. Disponible en: <http://www.bi-spain.com/articulo/70295/business-intelligence/otros/smart-process-management-spm-para-la-gestion-inteligente-de-incendios-forestales >

3. Detección de incendios forestales a través de imágenes digitales usando árboles de clasificación. Sistema desarrollado por Arturo Bustamante Blanco. [en línea]: Junio 2011. [Fecha de consulta: Julio 2011]. Disponible en:
   < http://perseo.cs.buap.mx/bellatrix/tesis/TES1411.pdf >

4. Análisis de variables sociodemográficas que inciden en incendios forestales, a través de técnicas de Data Mining. Sistema desarrollado por Hans Carlos Lucero Salinas. [en línea]: 2008 – 2011. [Fecha de consulta: Julio 2011]. Disponible en: <http://www.cifag.cl/_file/file_294_tesis_hans_lucero.pdf>

5. Aplicación de un Sistema de Información Geográfica al análisis de los datos de incendios forestales en España, elaborado por Rosa Almudena seco Granja. Sistema desarrollado por Rosa Almudena Seco Granja [en línea]: 2010 – 2011. [Fecha de consulta: Julio 2011]. <Disponible en: http://digital.csic.es/handle/10261/25971>

6. J. Hernández-Orallo, M. J. Ramírez-Quintana y C. Ferri. Introducción a la Minería de Datos. Prentice Hall / Addison-Wesley, 2008.

7. DataPrix, Introducción al Manual de DataWarehouse [en línea] [fecha de consulta: Noviembre 2011]. Disponible en: < http://www.dataprix.com/introduccion-manual-dwh >

8. Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. Second Edition. 2006

9. Sinnexus, Data mining(Minería de datos) [en línea][fecha de consulta: Noviembre 2011] Disponible en: < http://www.sinnexus.com/business_intelligence/datamining.aspx>

10. Pérez López César, Santín González, Daniel. Data mining. Soluciones con enterprise miner. Ra-Ma. 2006.

11. Nikhil R. Pal and Lakhmi Jain. Advanced Techniques in Knowledge Discovery and Data Mining. Springer. 2005

12. Sistema estatal de protección civil de Chiapas, Ciencias de la tierra para la sociedad, tema: incendios forestales [en línea]: Julio 2008 [fecha de consulta: Julio 2011] Disponible en: <http://www.proteccioncivil.chiapas.gob.mx/ciencia/ciencia&tierra/incendios.pdf >

13. Wikipedia, Microsoft SQL server [en línea]: 2011 [fecha de consulta: Julio 2011] Disponible en: <http://es.wikipedia.org/wiki/Microsoft_SQL_Server>

14. DataPrix, Data Warehouse vs Data Mart [en línea] [fecha de consulta: Noviembre 2011]. Disponible en: < http://www.dataprix.com/datawarehouse-vs-datamart>

15. Drivetoweb, Un poco de minería de datos [en línea] [fecha de consulta: Junio 2012]. Disponible en: < http://www.drivetoweb.com/?p=71>