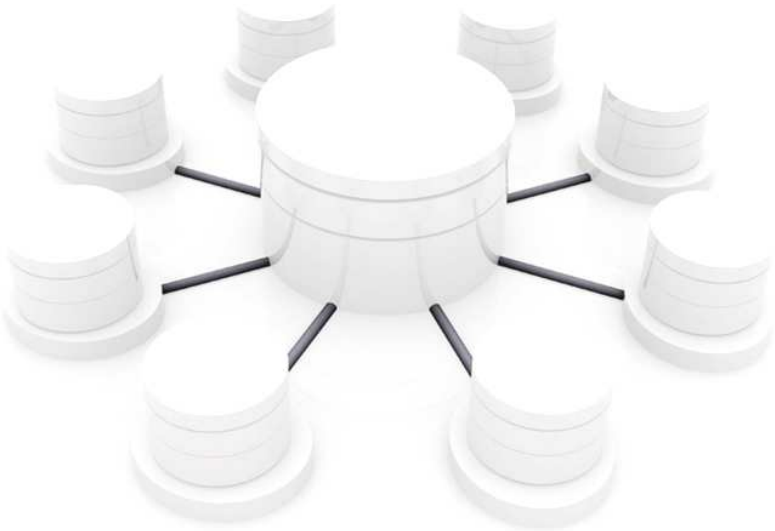
**BENÉMERITA UNIVERSIDAD AUTÓNOMA DE PUEBLA**
**FACULTAD DE CIENCIAS DE LA COMPUTACIÓN**

**TESIS PROFESIONAL**

QUE PARA OBTENER EL GRADO DE:
**LICENCIADO EN INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN**

# "SISTEMA PARA LA TOMA DE DECISIONES ACERCA DEL CAMBIO CLIMÁTICO"



PRESENTA
**Adriana Lopez Cumplido**

ASESOR:
**DRA. MARÍA JOSEFA SOMODEVILLA GARCÍA**

**PUEBLA, MÉXICO** **AGOSTO DE 2010**

# Agradecimientos

Agradezco…

……antes que nada a Dios por permitirme realizar todas las grandes o pequeñas cosas de las que se ha formado mi vida, y esta es una de ellas.

……..a mis padres, a quienes agradezco de todo corazón por su amor, cariño y comprensión. En todo momento los llevo conmigo.

……a mis hermanos por la compañía y el apoyo que me brindan. Sé que cuento con ellos siempre.

……a mi amor Alex, por tu apoyo, comprensión y amor que me permite sentir poder lograr lo que me proponga. Gracias por escucharme y tus concejos. Gracias por ser parte de mi vida, eres lo mejor….

…….a mí asesora, por haber confiado en mi persona, por la paciencia, sus consejos y su apoyo me ha conducido al buen término de este proyecto y que al final del camino existe un afecto muy grande hacia ella.

# Resumen

La percepción del Cambio Climático como uno de los problemas ambientales predominantes en el siglo XXI se ha venido reforzando en todo el mundo en los últimos años. Nuevas y crecientes evidencias del efecto de las interacciones del hombre con el medio ambiente se revelan ante nosotros con todo tipo de fenómenos irregulares que amenazan con cambiar bruscamente los patrones climáticos de la tierra, con efectos sin precedentes sobre los ecosistemas, la economía, la sociedad y para la propia sobrevivencia de la especie humana.

Con efectos de ampliar y profundizar en este problema es necesario conocer en qué medida contribuye la Ciudad de Puebla a la conformación del problema climático mediante sus emisiones de gases contaminantes, afines con las condiciones meteorológicas; es necesario responder cómo puede verse afectada por los impactos del Cambio Climático, qué acciones, políticas, medidas y estrategias se han venido adoptando para hacerle frente, cuáles son los escenarios y las perspectivas a corto y mediano plazo, cómo identificar oportunidades de cooperación ciudadana para enfrentar un desafío global que afecta a todos los humanos.

En los últimos años la Minería de Datos ha demostrado un auge para el soporte en la gestión del conocimiento. En este trabajo se presentan los resultados obtenidos al aplicar Técnicas de Minería de Datos a un conjunto de registros de mediciones de 3 contaminantes, Ozono, S02 y PM10, y 7 condiciones meteorológicas. Dichos registros fueron suministrados por la Secretaría del Medio Ambiente y Recursos Naturales del Estado de Puebla (SEMARNAT). Una de las técnicas aplicadas es agrupamiento (*K-Means*), la que indica el comportamiento de los contaminantes a lo largo del día en las diferentes estaciones del año. Otra técnica aplicada es la Regresión Lineal, que confirma que el ozono tiene una dependencia directa con la temperatura (ésta en mayor proporción), velocidad del viento, humedad relativa y los rayos UVA y UVB. Finalmente, se construyó un árbol de predicción (J48) con relación al Ozono para cada estación en función de la humedad, los rayos UV y otros. Estos hallazgos son de gran importancia en el rubro de Cambio Climático, el cual preocupa no sólo a gobiernos internacionales, nacionales y locales, sino a la sociedad en general. El resultado de este trabajo tiene un impacto en la predicción de contingencias ambientales en el área metropolitana de la Ciudad de Puebla, esto puede conllevar a una planeación estratégica para reaccionar ante contingencias, causadas por un contaminante del aire en particular el Ozono.

# Índice General


**1. Introducción** **8**
1.1 Planteamiento de la Investigación……..... 9
1.1.1 Problema a Resolver…….. 9
1.1.2 Objetivos de la Investigación…… 11
1.1.3 Justificación de la Investigación…… 12
1.2 Presentación de la Solución…… 13
1.2.1 Propuesta de solución…… 13
1.3 Aportaciones a la Investigación…… 13
1.4 Organización de la Tesis……..... 13
1.6 Conclusiones……..... 14
**2. Estado del Arte** **15**
2.1 Proyectos y esfuerzos para la mitigación del Cambio Climático…… 16
2.1.1 Proyecto CPTEC…….. 16
2.1.2 Proyecto CATHALAC…….. 17
2.1.3 Proyecto MACC…… 18
2.2 Trabajos recientes en Cambio Climático en México…… 19
2.3 El Cambio Climático y la Toma de Decisiones…… 20
2.4 Conclusiones…… 21
**3. Marco Teórico** **22**
3.1 ¿Qué es un Almacén de Datos…..…… 22
3.1.1 OLTP y OLAP…… 23
3.1.2 Almacenes de Datos y Bases de Datos Transaccionales…….. 24
3.1.3 Arquitectura de los almacenes de datos…… 25
3.1.3.1 Modelo Multidimensional…….. 25
3.1.3.2 Explotación de un almacén de datos. Operadores…….... 28
3.1.4 Carga y mantenimiento del almacén de datos…….. 29
3.1.5 Almacenes de datos y Minería de Datos…….. 31
3.2 Minería de Datos…….. 32
3.2.1 Introducción a la Minería de Datos…….. 33
3.2.1.1 Tipos de Tareas y Aplicaciones…… 33
3.2.2 Relación de la Minería de Datos con otras disciplinas…….. 35
3.2.3 La Minería de Datos y la Metodología KDD…… 36
3.2.3.1 Los Seis pasos del KDD y el proceso de Minería de Datos…… 37
3.2.4 Tareas y Técnicas de Minería de Datos…….. 39
3.2.4.1 Tareas Predictivas…….. 39
3.2.4.2 Tareas Descriptivas…… 41
3.2.5 Métodos. Correspondencia entre tareas y métodos…… 44
3.2.6 Técnicas de Minería de Datos Aplicadas …….. 46
3.2.6.1 *K-Means*…… 46
3.2.6.2 Regresión Lineal…… 49

3.2.6.3 Arboles de Decisión…….. 49
3.2.7 Entorno de Minería de Datos WEKA…… 51
3.2.8 Conclusiones…….. 53
**4. Análisis y Diseño** **54**
4.1 Planteamiento y requerimientos…….. 54
4.2 Fase de Integración y Recopilación…… 56
4.3 Fase de Selección, Limpieza y Transformación…… 58
4.4 Construcción del Almacén de Datos…… 59
4.5 Fase de Minería de Datos…….. 62
4.6 Conclusiones…….. 63
**5. Resultados** **64**
5.1 *Clustering*…… 64
5.2 Regresión Lineal…… 66
5.3 Árboles de decisión…… 67
5.4 Conclusiones…….. 68
**6. Conclusiones y Trabajo a Futuro** **69**
6.1 Aportaciones…….. 69
6.2 Líneas de Investigación Futuras…….. 70
6.3 Conclusiones Finales…… 70
**Referencias** **72**

# Índice de Figuras


| | | |
|---|---|---|
| Figura 2.1 | Modelo obtenido de Temperatura sobre Región de México (27 km) y Rep. Dominicana (9 km), por el proyecto CATHALAC [35]……..... | 18 |
| Figura 2.2 | Regiones impactadas por el ascenso en el nivel del mar…… | 19 |
| Figura 2.3 | Propuesta del Instituto de Investigación Internacional para el Clima y la Sociedad..... | 20 |
| Figura 3.1 | (a), (b). Visualización de un hecho en un modelo multidimensional……..... | 26 |
| Figura 3.2 | Modelo multidimensional…… | 28 |
| Figura 3.3. | Esquema de los procesos ETL……..... | 30 |
| Figura 3.4 | Perspectiva general y usos de un almacén de datos [2]…….. | 31 |
| Figura 3.5 | Disciplinas que contribuyen a la Minería de Datos [2]……..…….... | 35 |
| Figura 3.6 | Proceso de Extracción de Conocimiento [2]…….. | 38 |
| Figura 3.7 | Clústeres caracterizados por su centroide…….... | 47 |
| Figura 3.8 | Ejemplo de regresión lineal…… | 49 |
| Figura 3.9 | Ejemplo de Componentes y Estructura de un árbol de decisión…… | 50 |
| Figura 3.10 | Detalle del entorno Explorer de WEKA…… | 52 |
| Figura 4.1 | Fases del proceso de descubrimiento en bases de datos, KDD…… | 55 |
| Figura 4.2 | Ubicación de estaciones del Sistema Estatal de Monitoreo Atmosférico de Puebla [1]…….. | 57 |
| Figura 4.3 | Visualización de datos *outliers* con Weka……..... | 58 |
| Figura 4.4 | Ejemplo de discretización del atributo Fecha…… | 59 |
| Figura 4.5 | Modelo E-R de la base de datos ESTACIONES-SEMARNAT…… | 60 |
| Figura 4.6 | Dimensiones finales del cubo…… | 61 |
| Figura 4.7 | Vistas y Tablas…… | 61 |
| Figura 4.8 | Grupos de contaminación durante el día…… | 63 |
| Figura 5.1 | Muestra del árbol de decisión obtenido, con respecto de la estación 1……. | 67 |

# Índice de Tablas


| | | |
|---|---|---|
| Tabla 3.1. | Diferencias entre la base de datos transaccional y el almacén de datos…… | 25 |
| Tabla 3.2. | Correspondencia entre Tareas y Técnicas de la Minería de Datos [2]…… | 46 |
| Tabla 4.1 | Estaciones del Sistema Estatal de Monitoreo Atmosférico de Puebla……. | 56 |
| Tabla 4.2 | Parámetros del Sistema Estatal de Monitoreo Atmosférico de Puebla……. | 56 |
| Tabla 5.1. | Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 1…….. | 64 |
| Tabla 5.2. | Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 2…….. | 65 |
| Tabla 5.3. | Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 3…….. | 65 |
| Tabla 5.4. | Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 4…….. | 65 |
| Tabla 5.5. | Ecuaciones para predecir los niveles de Ozono de las principales mediciones meteorológicas…….. | 66 |
| Tabla 5.6. | Comparativo entre los niveles de Ozono reales y las predicciones…….. | 67 |

# Capítulo 1

# Introducción

Actualmente, el tema de Cambio Climático se ha convertido en el asunto ambiental más complejo visto en la agenda de la política internacional y nacional, ya que es uno de los problemas más difíciles de tratar, representa desafíos a la sociedad como un todo, a la comunidad científica y técnica y a las autoridades políticas. Generalmente se acepta que el calentamiento global del planeta se debe al desequilibrio que la actividad humana ha provocado en las concentraciones atmosféricas de los gases de efecto invernadero.

Ante este nuevo entorno, se dispone de información valiosa que ayuda afijar estrategias y tomar decisiones, la información se contempla desde esta perspectiva, como el recurso vital del Medio Ambiente, que precisa ser integrada para que realmente añada valor e innovación.

En lo expuesto anteriormente, la información y el conocimiento, se convierten en dos factores clave para aumentar el manejo inteligente en la toma de decisiones de las cuales, se obtendrá información mucho más detallada y variable en periodos de tiempo. De las herramientas que actualmente ofrece el mercado consideramos los *DataWarehouse* (Almacenes de Datos), que son tecnologías idóneas para las organizaciones.

Los *DataWarehouse* (DW), son el centro de atención hoy en día de las grandes organizaciones, ya que componen una de las columnas elementales para el proceso de toma de decisiones, de aquí que la información depositada en ellos sea confiable y de calidad. Dada la competencia existente, que crece en todo momento, estas decisiones deben ser rápidas y ser tomadas sobre una gran cantidad de hechos y cifras. En concreto necesitamos herramientas que nos ayuden a minimizar el tiempo para analizar toda esa información con mayor velocidad y precisión; logrando de esta manera mantenernos competitivos, y reaccionar ante cambios potenciales que afecten la vida del ser humano.

En este trabajo se aborda la construcción de un *DataWarehouse* con respecto al actual Cambio Climático, para su posterior explotación a través de Técnicas de Minería de Datos. La información de base para este proyecto fue suministrada por la Secretaría del Medio Ambiente y Recursos Naturales del Estado de Puebla. Dicha información se encuentra almacenada en bases de datos digitales y otras fuentes. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Este proyecto constituye todo un desafío para la Minería de Datos, ya que ello implica trabajar con un volumen de datos considerable y esto conlleva a problemas presentes sobre los datos como: ruido, datos ausentes, intratabilidad, volatilidad de los datos, entre otros. Se precisa conjuntar dicha información, y para ello se propone como solución la intervención de los almacenes de datos, las operaciones OLAP (*Online Analitical Processing*, por sus siglas en inglés), y técnicas adecuadas de Minería de Datos para analizar y extraer conocimiento novedoso y útil. La aplicación de dichas herramientas facilita el problema de acceso a la información y en consecuencia, se acelera el proceso de análisis, consultas y el menor tiempo posible de uso de la información.

## 1.1 Planteamiento de la Investigación

En esta sección se precisa el problema de la investigación a resolver, se definen los objetivos del proyecto, al mismo tiempo se plantea la propuesta de solución y por último se describe la organización de la tesis.

### 1.1.1 Problema a Resolver

El problema que se aborda en este proyecto tiene sus orígenes en datos otorgados por la Secretaria del Medio Ambiente y Recursos del Estado de Puebla (SEMARNAT).

En junio de 2000 la Secretaría de Desarrollo Urbano, Ecología y Obra Pública, SEDURBECOP, instaló e inició operaciones de una red de calidad del aire con cuatro estaciones de monitoreo automático, que incluyen la medición de contaminantes como Ozono, óxidos de nitrógeno, dióxido de azufre, ácido sulfhídrico, hidrocarburos, $PM_{10}$ y meteorología, esta última está conformada por los siguientes parámetros: temperatura, humedad relativa, radiación UV-A , radiación UV-B, velocidad del viento, dirección del viento, presión barométrica. Actualmente, el Sistema Estatal de Monitoreo Ambiental, SEMA, se encuentra a cargo de la SEMARNAT, del gobierno del Estado de Puebla.

Con fundamento en los resultados de la medición de la calidad del aire, se han podido desarrollar varios proyectos como los siguientes: Programa de Gestión de la Calidad del Aire en la Zona Metropolitana del Valle de Puebla 2006-2011, inventario de emisiones 2004, estudios de mitigación de Puebla ante el Cambio Climático, modelación de la calidad del aire, modernización del programa de verificación vehicular, mecanismo de desarrollo limpio, control del gas metano del municipio de Puebla y estudios epidemiológicos [1].

Este proyecto igualmente se cimenta en las mediciones obtenidas de la calidad del aire, sin embargo la información que recaba cada una de estas estaciones, presenta problemas importantes, mismos que son mencionados a continuación:

- Los datos proceden de fuentes diversas y pertenecen a diferentes dominios. Aquí es clara la necesidad de integración y considerar los mismos para la obtención de información útil para la organización.
- En la actualidad se utiliza un método tradicional de convertir los datos en conocimiento, el cual consiste en un análisis e interpretación realizados de forma manual. El especialista en la materia, digamos por ejemplo un Ingeniero Ambiental analiza los datos y elabora un informe o hipótesis que refleja las tendencias o pautas de los mismos. Esta forma de operar es pesada, cara y altamente subjetiva.
- En consecuencia del punto anterior, muchas decisiones importantes se realizan, no sobre la base de la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias.
- La información recabada por cada una de las estaciones representa trabajar con grandes volúmenes de datos, donde la enorme abundancia de datos desborda la capacidad humana de comprenderlos.
- La información recabada conlleva problemas como ruido, datos ausentes, intratabilidad, volatilidad de los datos…etc.

El objetivo de este proyecto es ambicioso pero factible, ya que el objetivo de la Minería de Datos es convertir datos en conocimiento.

## 1.1.2 Objetivos de la Investigación

El objetivo general de la tesis es el siguiente:

***Diseñar una Data Warehouse para la toma de decisiones con respecto al Cambio Climático actual, y posteriormente explotarlo con Técnicas adecuadas de Minería de Datos.***

Los objetivos particulares se especifican a continuación:

1. Permitir la generación de escenarios futuros del clima, con respecto al efecto de los contaminantes, en las diferentes estaciones del año.
2. Realizar búsquedas de comportamiento de contaminantes en relación a ciertas variables.
3. Realizar búsquedas de cambios en el comportamiento de contaminantes a diferentes horas del día.
4. Facilitar el lanzamiento de campañas organizacionales más adecuadas para mitigar el Cambio Climático.
5. Mejorar el análisis de riesgos. Con esto se plantea también incrementar la confiabilidad de la información generada por la SEMARNAT.
6. Facilitar el proceso de toma de decisiones basado en información estadística.
7. Identificar las patologías para el diagnóstico de ciertas enfermedades relacionadas con las partículas $PM_{10}$[1].
8. Permitir la detección de ciertos pacientes con riesgo de sufrir una patología concreta.
9. Contribuir a una mejor adaptación del ser humano al actual calentamiento global.

## 1.1.3 Justificación de la Investigación

Los siguientes criterios justifican la investigación en la que se fundamenta este trabajo de tesis:

1. **Reducción de incertidumbre sobre aspectos de la realidad**. El actual Cambio Climático es un problema con características únicas, ya que es de naturaleza global, sus impactos mayores serán a largo plazo e involucra interacciones complejas entre procesos naturales, sociales, económico y políticos [4]. Los problemas y las dudas a la hora de cubrir estas necesidades aparecen si se desconoce por dónde empezar, qué herramientas utilizar, qué técnicas estadísticas o de aprendizaje automático son más apropiadas y qué tipo de conocimiento se puede llegar a obtener y con qué fiabilidad. En este trabajo se intentan resolver estas dudas, pero además, se presentan una serie de posibilidades y, por supuesto, de limitaciones.

2. **La necesidad e impulso creciente del área de Minería de Datos**. Las tecnologías basadas en la informática, Meteorología, y Cambio Climático, son claves para este desarrollo, pues en ellas se suministran potentes instrumentos para la obtención y el análisis de la información climatológica. La aparición de nuevas tecnologías ha posibilitado el proceso de investigación respecto del Cambio Climático, al facilitar el estudio de las interacciones de los contaminantes y su influencia en el desarrollo de enfermedades.

3. En las últimas décadas, el almacenamiento, organización y recuperación de la información se ha automatizado gracias a los sistemas de bases de datos, así como para el descubrimiento del significado que poseen los datos almacenados en grandes bancos. Esto permite explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones; además de facilitar la extracción de la información existente en los textos, así como para crear sistemas inteligentes capaces de entenderlos.

4. Todos estos problemas y limitaciones de las aproximaciones clásicas han hecho surgir la necesidad de una *nueva generación de herramientas y técnicas* para soportar la extracción de conocimiento útil desde la información disponible, y que se engloban bajo la denominación de Minería de Datos.

## 1.2 Presentación de la Solución

La propuesta de solución al problema definido en la sección 1.1.1 y los productos desarrollados son comentados en la siguiente subsección.

## 1.2.1 Propuesta de solución

De acuerdo con el problema planteado, se presenta la siguiente definición conceptual de la solución:

"Se propone el diseño, construcción e implantación de un Sistema de Data Warehouse Multidimensional para la toma de decisiones, respecto al actual Cambio Climático. Una vez creado el *DataWarehouse* se explotará con Técnicas de Minería de Datos adecuadas".

El sistema propuesto se fundamenta en la Teoría de Bases de Datos Relacionales y Construcción de Cubos en repositorios de Datos a partir de SQL *Server Business Intelligence Development Studio*, así como también el uso de Weka (Waikato *Environment for Knowledge Analysis*-Entorno para Análisis del Conocimiento de la Universidad de Waikato), un conocido software para aprendizaje automático y Minería de Datos escrito en Java y desarrollado por la Universidad de Waikato. Weka es un software de distribución libre, bajo licencia de GNU-GPL[1].

## 1.3 Aportaciones a la Investigación

Las aportaciones se derivan de la comparación de las técnicas de Minería de Datos, aplicadas sobre el *DataWarehouse*, enriqueciéndose por medio de la visualización de ejemplos, y de los alcances de trabajos similares presentados en el Capitulo 2, "Estado del Arte".

## 1.4 Organización de la Tesis

Este trabajo se estructura en 6 capítulos distribuidos de la siguiente manera:

- Capítulo I. Introducción. En esta parte se detalla el problema a resolver, los objetivos de la investigación, la justificación, así como también se presenta la propuesta de solución al problema y los productos generados por el trabajo de la presente tesis. Al mismo

[1] Disponible en : http://es.wikipedia.org/wiki/Licencia_pública_general_de_GNU

tiempo, se presentan las aportaciones que genera la explotación del almacén de datos y las Técnicas de Minería de Datos, así mismo se exhibe la organización del documento.

- Capítulo II. Estado del Arte. En esta parte se detallan un número importante de trabajos de investigación relacionados con el ámbito de los *DataWarehouse* y la Minería de Datos.
- Capítulo III. Marco Teórico. En este capítulo se profundiza acerca del conocimiento disponible acerca de los Almacenes de Datos y las Técnicas de Minería de Datos. Al mismo tiempo se profundiza en los fundamentos teóricos que ofrecen el sustento formal al desarrollo de la tesis.
- Capítulo IV. Análisis y Diseño de la construcción de un *DataWarehouse* para la toma de decisiones con respecto del Cambio Climático. En este capítulo se presenta el análisis del problema planteado en la sección 1.1 y se muestra el proceso de diseño del mismo.
- Capitulo V. Implementación y Resultados. En este capítulo se presentan los resultados obtenidos de la aplicación de Técnicas de Minería de Datos al Diseño del *DataWarehouse* presentado en el capítulo IV.
- Capítulo VI. Conclusiones y Trabajo a Futuro. En esta parte se detallan las conclusiones de la investigación a través de tres secciones dedicadas a exponer una discusión sobre las lecciones aprendidas durante el desarrollo de la tesis.
- Finalmente se da una lista de referencias y material consultado.

## 1.5 Conclusiones

El Cambio Climático es una amenaza para la humanidad, pero nadie puede determinar con seguridad sus futuros efectos o la magnitud de éstos. En casi todos los círculos científicos se trata de ver en qué forma se solucionará y cuál será la mejor forma de detectar las repercusiones. Las tecnologías del *DataWarehouse* y Minería de datos vienen desempeñando un papel cada vez más importante en cuanto al soporte para el análisis del Cambio Climático, es por ello que en este capítulo se presentó el protocolo de investigación para desarrollar el sistema de toma de decisiones acerca del Cambio Climático.

# Capítulo 2

# Estado del Arte

El problema del Cambio Climático ha originado un número importante de trabajos de investigación durante los últimos años en el ámbito de los *DataWarehouse* y la Minería de Datos. Haciendo una revisión a la literatura se resumen algunas de las investigaciones que se han realizado sobre los impactos actuales y potenciales del Cambio Climático, recurriendo a la tecnología de *DataWarehouse* y Minería de Datos.

La referencia de esta área de investigación es del artículo "*Climate Change, Vulnerability and Adaptation in Latin Americ*a", este introduce de una manera unificada y sistemática los problemas relacionados con el Cambio Climático en la Latinoamérica actual, así como también muestra las medidas que se han adoptado y algunos proyectos, universidades e instituciones del clima, que se encuentran trabajando sobre este problema. El artículo "*Change on British Columbia's Biodiversity*", aporta información complementaria sobre el tema de investigación.

En la actualidad existen condiciones evidentes de un Cambio Climático inevitable, con características irrevocables como: calentamiento global, deterioro universal de los *hábitats* naturales y medios de subsistencia. Estos efectos secundarios también se extienden a nivel del mar, disminución de suministro de agua dulce, impactos sobre la agricultura y la salud de América Latina.

A partir de este punto nos concentraremos en los trabajos realizados por diversas instituciones, resaltando el uso de *DataWarehouse* y la Minería de Datos, para los cuales presentaremos un estado del arte detallado.

## 2.1 Proyectos y Esfuerzos para la mitigación del Cambio Climático en Latinoamérica

Muchas universidades, instituciones de investigación del clima y organizaciones no gubernamentales, han realizado un esfuerzo por promover modelos climáticos regionales y escenarios del clima. Algunos ejemplos de estos trabajos se mencionan en las siguientes secciones.

### 2.1.1 Proyecto CPTEC

El Centro de Previsión del Tiempo y Estudios Climáticos **(CPTEC)** en Brasil, es un centro líder en investigación para el sistema climático, análisis, modelización y predicción del clima. Este centro tiene como objetivo el desarrollo de investigación sobre el tema de Cambio Climático, incluyendo estudios de observación para caracterizar el clima actual y su variabilidad a largo plazo, así como estudios de proyecciones de escenarios climáticos futuros para caracterizar el clima del resto del siglo XXI, para diferentes escenarios de emisiones de gases de efecto invernadero. Entre los miembros, hay investigadores que trabajan en áreas de Cambio Climático, análisis de vulnerabilidad, estudios de impacto y adaptación. Las instituciones que participan son de la talla de la Universidad de Sao Pablo de la Fundación Brasileña para el Desarrollo sostenible, en colaboración con las instituciones del Gobierno Federal Brasileño (EMBRAPA, INMET, FIOCRUZ, ANA, ANEEL, ONS, COPPE-UFRJ, entre otros), así como centros meteorológicos estatales, universidades, la FBMC (Foro Brasileño de Cambio Climático) y organizaciones de la sociedad civil. El grupo también trabaja en estrecha colaboración con el Consejo Nacional del Clima conjunto al Departamento de Cambio Climático y la Calidad del Aire del Ministerio de Medio Ambiente [34].

El trabajo que se desarrolla es **proporcionar información y pronósticos** del tiempo para darse a conocer y puestas a disposición por los grupos del clima y la investigación aplicada, para dar apoyo a la **toma de decisiones** en la formulación de políticas sobre el impacto del Cambio Climático [34].

Algunos de los trabajos relacionados con esta organización son los siguientes:

- Impacto sobre la salud humana de las partículas emitidas por quemas en la Amazonia Brasileña. Ignotti, E., Valente, J., Longo, Karla M. Freitas, SR, Hacon, S., Artaxo, P. Revista de Salud Pública.
- Aire Nuevo producto de Calidad en el CPTEC/INPE: Pronósticos y de sus precursores de ozono troposférico de la quema de biomasa y emisiones urbanas. KM Longo, Freitas, SR, Alonso, M., Rodrigues, LF. , Mello, R., Stockler, R, Moreira, D.
- Entre otras.

## 2.1.2 Proyecto CATHALAC

El proyecto **CATHALAC** (Organismo Internacional dedicado a promover el desarrollo sostenible en América Latina y el Caribe por medio de la investigación aplicada, la educación y la transferencia de tecnología), trabaja con un *DataWarehouse*, y ayuda a la evaluación de la capacidad de Adaptación en América Central, México y Cuba. Desde 2005 el proyecto CATHALAC, incluye el terminó de *Clusters*, para ayudar al procesamiento de datos referentes a modelación del clima y de Cambio Climático, esto dentro del marco del proyecto SERVIR (*The Decision Support System For Environmental Management* – según sus siglas en Inglés). En el proyecto CATHALAC se cuenta con 2 CLUSTERS en los cuales corren los modelos MM5 y WRF que son utilizados para pronóstico del clima para periodos cortos (2 a 3 días de pronóstico) y PRECIS para modelación de cambio Climático (20 a 100 años de Cambio Climático). El proceso inicia en la madrugada donde automáticamente se descargan los datos de entrada de los modelos para después ser analizados y procesados por el CLUSTER1 que consta de 9 nodos esclavos y 1 maestro. Luego de la primera ronda de modelación de trabajos se inicia con la Región de México (27 km) y Rep. Dominicana (9 km) (Figura 2.1), esta segunda ejecución utiliza los datos descargados de la primera ejecución y alimenta al modelo. En CATHALAC se encontraba en ejecución el modelo de PRECIS que demoró 5.1 meses para 100 años de modelado de Cambio Climático. En los próximos meses se espera tener nuevos modelos climáticos funcionando (ver figura 2.1) y se comenzará con un sistema de CLUSTERS con lo que cuenta el centro, para que CATHALAC se posicione como pionero en supercómputo en la región.

## 2.1.3 Proyecto MACC

El proyecto **MACC** (*Monitoring Atmospheric Composition and Climate* – por sus siglas en Inglés) es financiado por el Fondo para el Medio Ambiente Mundial (FMAM), es un proyecto regional llevado a cabo en doce países de la CARICOM (*Caribbean Community* –por sus siglas en Inglés). El objetivo principal es seguir fortaleciendo la capacidad de los pequeños Estados Insulares de bajo desarrollo y ayudar a los Estados Ribereños del Caribe para aumentar su resistencia al Cambio Climático, mediante la identificación y aplicación de las opciones viables de adaptación. MACC ofrece registros de datos sobre la composición atmosférica de los últimos años, los datos de seguimiento de las condiciones actuales y las previsiones sobre la distribución de los componentes clave para unos días antes. MACC combina la modelización atmosférica, utilizando esta técnica recaban datos de observación de la Tierra para prestar servicios información, que abarcan la calidad del aire Europeo, Composición de la Atmósfera Global, el clima, los rayos UV y la energía solar [33].

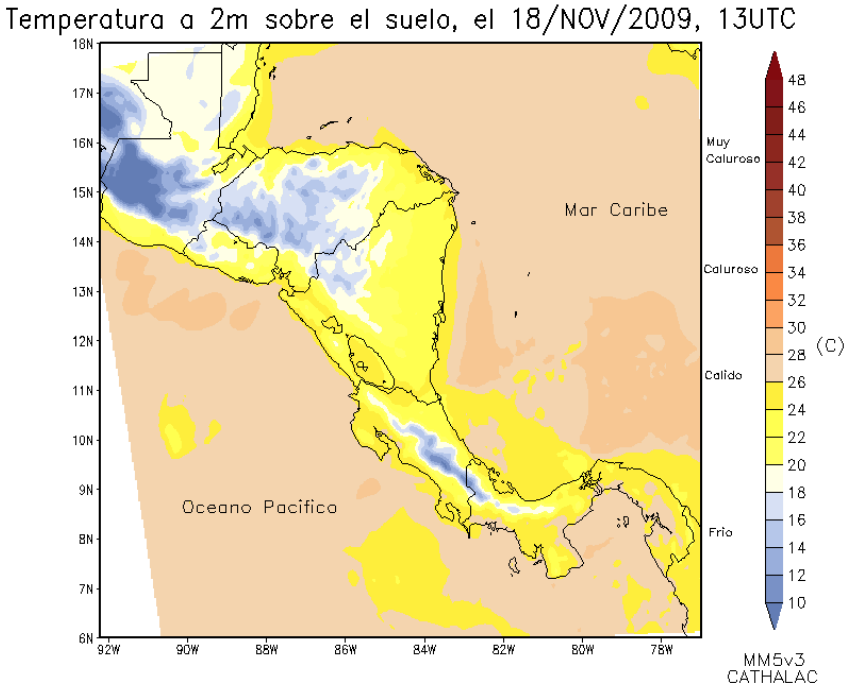



Figura 2.1: Modelo obtenido de Temperatura sobre Región de México (27 km) y Rep. Dominicana (9 km), por el proyecto CATHALAC [35].

## 2.2 Trabajos recientes en Cambio Climático en México

Recientemente apareció la revisión del estado actual del conocimiento sobre Cambio Climático elaborada por el Panel Intergubernamental para el estudio del Cambio Climático, conocido como IPCC por sus siglas en inglés. En esta revisión se analiza básicamente lo relacionado con:

- El inventario nacional de emisiones
- La vulnerabilidad ante el Cambio Climático
- Medidas de mitigación

Con respecto a los análisis sobre el aumento del calentamiento en otras partes de México ya se comienza a reflejar en un aumento en el número de eventos extremos. Las evidencias observacionales en ese sentido no permiten hasta ahora concluir al respecto.

El calentamiento global viene acompañado por una elevación del nivel del mar (ver Figura 2.2) debido a la expansión térmica de los océanos; ésta se traduce en que zonas costeras bajas- por ejemplo por debajo de los 2 m por arriba de la marea alta- se vuelven vulnerables a las inundaciones. En la figura siguiente se ilustra cómo Tamaulipas, Veracruz, Tabasco, Yucatán y Quintana Roo se verían afectados [37].

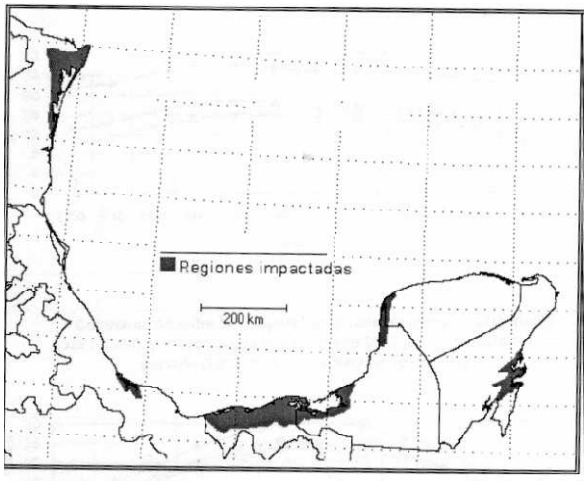



Figura 2.2: Regiones impactadas por el ascenso en el nivel del mar.

Un elemento importante en materia de escenarios de Cambio Climático es la selección objetiva de los modelos de circulación general que se utilizan para generar escenarios futuros de las variables climáticas. Por ejemplo, para el caso de la temperatura, se encuentra que los modelos HADCM2 y EHCAM4 generan los mejores escenarios de cambio de temperatura en México. Las gráficas siguientes muestran una comparación del ciclo anual de la temperatura pronosticada para el año 2050 y de la temperatura observada en algunas regiones de México.

## 2.3 El Cambio Climático y la Toma de Decisiones

Los decisores o tomadores de decisión (incluyendo responsables en formulación de políticas) que trabajan en los sectores públicos y privativos de países en vías de desarrollo, típicamente enfrentan la presión de actuar en respuesta a problemas que requieren acción inmediata. Además, el efecto de tales decisiones debe ser evidente durante plazos, usualmente cortos, en los cuales esos decisores operan. Consecuentemente, dan prioridad relativamente baja a los asuntos que se perciben como problemas de un futuro distante, tal es el caso del "Cambio Climático". En el Instituto de Investigación Internacional para el Clima y la Sociedad (IRI), se propone de manera efectiva ayudar a las sociedades a estar preparadas y adaptadas para cualquier escenario posible de Cambio Climático (Figura 2.3). Esto requiere establecer una evaluación de riesgo climático y estrategias de manejo de riesgo [36].


Figura 2.3: Propuesta del Instituto de Investigación Internacional para el Clima y la Sociedad.

## 2.4 Conclusiones

En este capítulo se presentaron proyectos relacionados estrechamente con el Cambio Climático, los *DataWarehouse* y la Minería de datos, en Latinoamérica. Se trató además, sobre la Iniciativa en la toma de decisiones con respecto del actual Cambio Climático.

# Capítulo 3

# Marco Teórico

En este capítulo se presentan los fundamentos teóricos que ofrecen el sustento formal al desarrollo de la tesis. Exponiendo como primer punto la tecnología de los *DataWarehouse* (DW) como una herramienta para analizar la información y como segundo punto, además se expondrá la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados. Con respecto al segundo punto que involucra la Minería de datos, cabe mencionar que es la parte más extensa de este trabajo de tesis y ligeramente la más técnica, ya que se trata de describir el funcionamiento y la conveniencia de las técnicas de Minería de Datos aplicadas.

## 3.1 ¿Qué es un Almacén de Datos?

Desde el inicio las bases de datos se han transformado en un instrumento primordial de control y conducción de operaciones comerciales. Fue así como en breves años en las grandes empresas y negocios coexistía un considerable número de información almacenada en diferentes fuentes de datos y éstas alcanzaron un tamaño considerablemente grande. Con este gran cúmulo de información, los dirigentes de tales empresas y dependencias se dieron cuenta que podría tener un resultado útil, al estar exhibida la suma de sus operaciones comerciales. Por tanto, se preocuparon por unificar las diversas fuentes de información de las cuales disponían, en un único lugar, al que sólo se le agregaría información importante, sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. La respuesta a esta problemática fueron los Almacenes de Datos o *DataWarehouse*.

Existen muchas definiciones para el Almacén de Datos, la más conocida fue propuesta por William Inmon- considerado el padre del *DataWarehouse* en 1992 [6]:

*"Un DW es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales". William Inmon indicó que un DataWarehouse se caracterizaba por ser:*

- **Temático**: Los DW están diseñados para ayudar a analizar los datos de un determinado tema o significado. Por ejemplo en este proyecto se desea saber más sobre los efectos que produce el Ozono sobre el Cambio Climático, sobre estas bases se puede construir un DW, que concentre consultas. Utilizando este DW se podrían hacer peguntas del tipo ¿Existe una relación entre el Ozono y los diferentes contaminantes?, que puedan determinar la toma de ciertas decisiones con respecto de salud, para beneficio de la población Mexicana. Está habilidad de localizar un tema prioritario hace que se cree un DW orientado a un tema. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

- **Integrado**: La integración está muy relacionada con el punto "temático". Los DW deben aunar datos de fuentes dispares de una forma consistente. Deben resolver problemas tales como el nombre de los campos, conflictos de inconsistencia en unidades y medidas antes de ser almacenados.

- **Variante en el tiempo:** Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones. Para esto se necesita una gran cantidad de datos almacenados a lo largo de mucho tiempo. En esto difiere mucho un sistema transaccional, donde los datos históricos son archivados y poco accedidos. La información del depósito por el contraste, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias.

- **No volátil:** La información es útil sólo cuando es estable [8]. La información no se modifica ni se elimina, una vez almacenado el dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas. Es lógico debido a que el designio del *DataWarehouse* es ser capaz de analizar lo que ya ha sucedido.

### 3.1.1 OLTP y OLAP

Existen dos tipos de sistemas orientadas a los procesamientos muy diferentes denominados OLAP Y OLTP los cuales se detallan a continuación:

- OLTP (*On-Line Transactional Processing*). Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones en tiempo real, constituye el trabajo primario en un sistema de información. Este trabajo consiste en realizar transacciones, es decir, actualizaciones y consultas a la base de datos con un objetivo operacional [2].

- OLAP (*On-Line Analytical Processing*). Son bases de datos orientadas al procesamiento analítico en tiempo real, que engloba un conjunto de operaciones, exclusivamente de consulta, en las que se requiere agregar y cruzar gran cantidad de información. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. Ejemplos de este tipo de trabajo analítico pueden ser: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos… etc.

Una característica de ambos procesamientos es que se plantea 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 está tan claro que esto sea posible para algunas consultas muy complejas realizadas por el OLAP [2].

### 3.1.2 Almacenes de Datos y Bases de Datos Transaccionales

Tradicionalmente el análisis para la toma de decisiones se realizaba sobre las mismas bases de datos de trabajo o base de datos transaccionales. Esto implica combinar el trabajo transaccional diario de los sistemas de información originales (OLTP), con el análisis de datos en tiempo real sobre la misma base de datos (OLAP), esto provoca problemas como que:

- Disturba el trabajo transaccional diario de los sistemas de información originales.
- Se realizan consultas muy pesadas (*killer queries*).
- En situaciones de carga alta, la perturbación es tal que el proceso analítico se debe realizar por la noche o en periodos festivos.

Todos estos problemas son provocados porque la base de datos está diseñada para el trabajo transaccional y no para el análisis de los datos, por lo que el análisis es lento.

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, entonces:

- Facilita el análisis de los datos en tiempo real (OLAP).

- No disturba el OLTP de las base de datos originales.

Es por ello que no debemos confundir y por lo tanto diferenciar claramente entre las bases de datos transaccionales (u operacionales) y los almacenes de datos. De hecho hoy en día las diferencias son claras, como se muestran en la tabla 3.1.

Tabla 3.1. Diferencias 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… |
| **Modelo de datos** | Datos normalizados. | Datos en estrella, en copo de nieve, parcialmente desnormalizados, multidimensionales… |
| **Número y tipo de usuarios** | Cientos/miles: aplicaciones, operarios, administrados 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. |

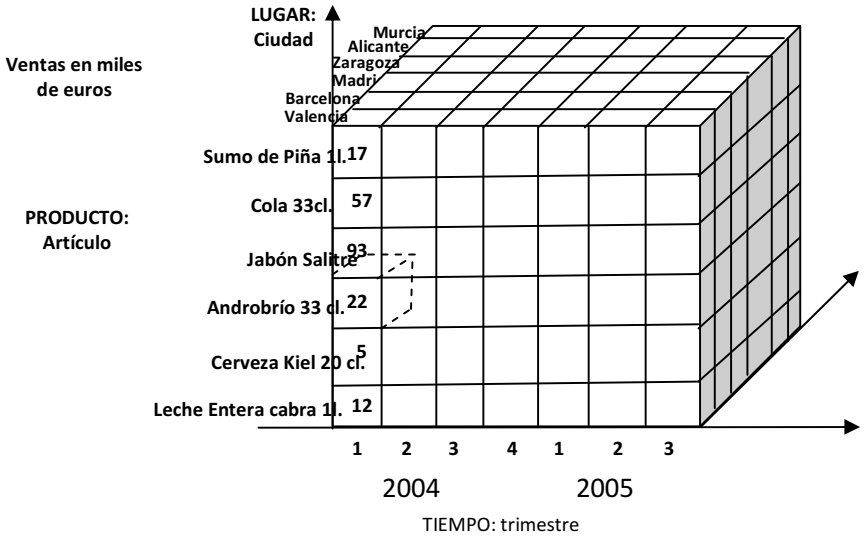
## 3.1.3 Arquitectura de los almacenes de datos

Un almacén de datos recaba, principalmente, datos históricos, es decir, hechos, sobre el contexto en el que se desenvuelve la organización. Los hechos son, por tanto, el aspecto central de los almacenes de datos. Esta característica determina en gran medida la manera de organizar los almacenes de datos.

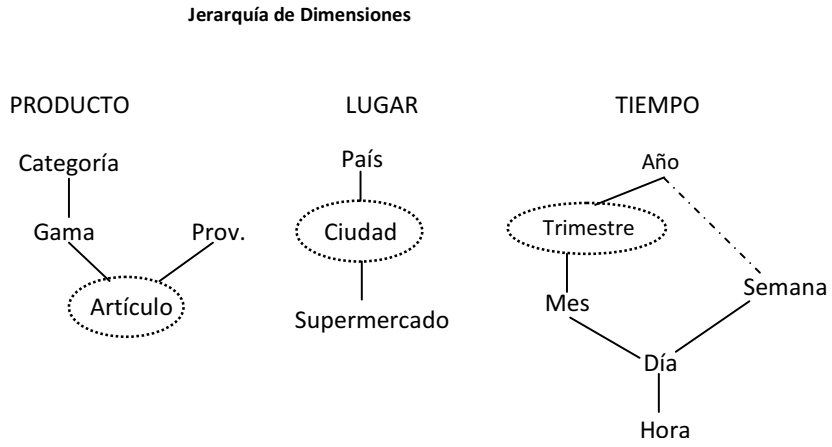
## 3.1.3.1 Modelo Multidimensional

El modelo conceptual de datos más extendido para los almacenes de datos es el modelo multidimensional [2] Ver Figura 2.3. (a). El modelo multidimensional representa los datos como matrices N-Dimensionales denominadas Hipercubos (*data cubes*-cubos de datos). Un hipercubo consiste en un conjunto de celdas, cada una se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones [7]. Este modelado de datos forma conjuntos como medidas descritas por dimensiones, las cuales son adecuadas para resumir y organizar datos (p.ej. hojas de cálculo), también está enfocado para trabajar sobre datos de tipo numérico, los cuales son mucho más simples de entender y fáciles de visualizar que el modelado Entidad Relación. Los

datos en el hipercubo se organizan en torno a los *hechos*, que tienen unos atributos o *medidas* que pueden verse en mayor o menor detalle según ciertas *dimensiones* [2].




*a)* **Modelo Multidimensional**




**b) Jerarquía de Dimensiones**

Figura 3.1: (a), (b). Visualización de un hecho en un modelo multidimensional.

En la Figura 3.1. (a) se representa un cubo tridimensional donde las dimensiones producto, lugar y tiempo se han agregado por artículo, ciudad y trimestre. La representación de

un hecho como el visto anteriormente corresponde, por tanto, a una casilla en dicho cubo. El valor de la casilla es la medida observada (en este caso el importe de las ventas). Esta visualización hace que, incluso cuando tengamos más de tres dimensiones, se habla de un "cubo" (o más propiamente de un "hipercubo") como un conjunto de niveles de agregación para todas las dimensiones. Esta estructura permite ver de una manera intuitiva la sumarización/agregación (varias casillas se fusionan en casillas más grandes), la disgregación (las casillas se separan en casillas con mayor detalle) y la navegación según las dimensiones de la estrella.

A continuación, se describen cada uno de los componentes de la organización de un hipercubo:

- **Hecho:** es el objeto a analizar, posee atributos llamados de hecho o de síntesis, y son de tipo cuantitativo. Sus valores (medidas) se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor. Por ejemplo: cantidad de unidades en inventario, ventas en dólares, cantidad en unidades de producto vendidas, horas trabajadas, promedio de piezas producidas, consumo de combustible de un vehículo, etcétera [7].

- **Dimensiones:** representan cada uno de los ejes en un espacio multidimensional. Suministran el contexto en el que se obtienen las medidas de un hecho. Algunos ejemplos son: tiempo, producto, cliente, departamento, entre otras. Las dimensiones son entidades o perspectivas que sirven para mantener estructurados los datos de una organización, además se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Los componentes de una dimensión se denominan *niveles* y representan nombres o identificadores que marcan una posición dentro de la dimensión y se agrupan de forma jerárquica (p.ej. Meses, trimestres y años son miembros de la dimensión tiempo, ciudades, regiones y países son miembros de la dimensión localización.), los miembros de las dimensiones se suelen organizar en forma de jerarquías (Figura 3.1 (b)).

Los hechos se guardan en tablas de hechos y las dimensiones en tablas de dimensiones.

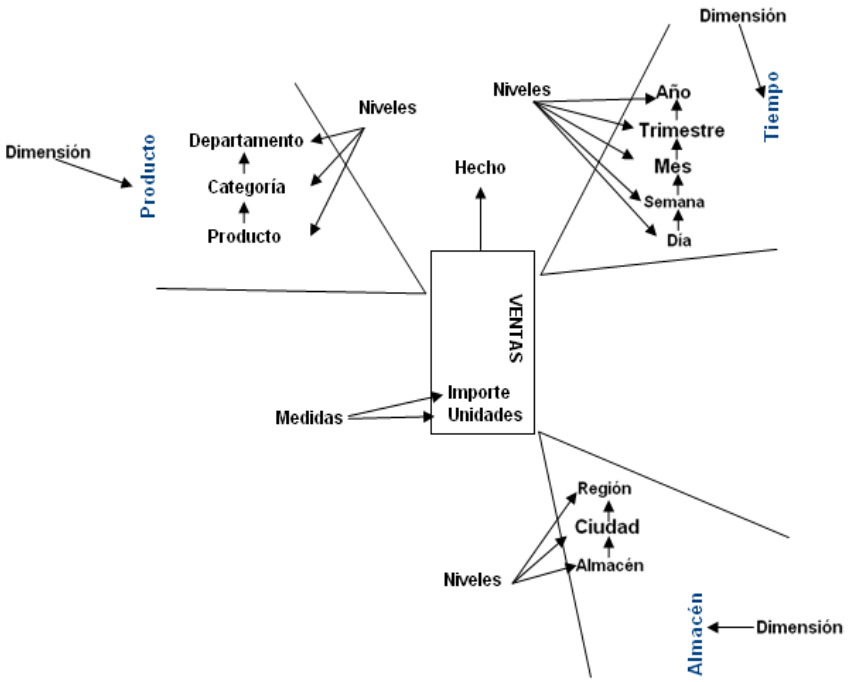



Figura 3.2: Modelo Multidimensional [7].

En la figura 2.4, se muestra un modelo multidimensional, donde los hechos se encuentran en una tabla de ventas y las dimensiones son almacén, producto y tiempo. Un modelo multidimensional se puede representar como un esquema en estrella, copo de nieve (*snowflake*) o constelación de hechos [9].

## 3.1.3.2 Explotación de un almacén de datos. Operadores

Lo interesante de un Almacén de Datos no es poder realizar consultas que, en cierto modo, se pueden hacer con selecciones, proyecciones, concatenaciones y agrupamientos tradicionales. Lo realmente interesante de las herramientas OLAP son sus operadores de refinamiento o manipulación de consultas. El carácter agregado de las consultas en el Análisis de Datos, aconseja la definición de nuevos operadores que faciliten la agregación (consolidación) y la disgregación (división) de los datos.

Los operadores más importantes asociados al modelo multidimensional se detallan a continuación:

- ***Drill*** (disgregación). Se trata de disgregar los datos (mayor nivel de detalle o desglose, menos sumarización) siguiendo los caminos de una o más dimensiones [2].
- ***Roll*** **(**agregación**).** Se trata de agregar los datos (menor nivel de detalle o desglose, más sumarización o consolidación) siguiendo los caminos de una o más dimensiones [2].
- ***Slice & Dice*****.** Se seleccionan o proyectan los datos
- ***Pivot*****.** Se reorientan las dimensiones.

A la acción de navegar por los datos del *DataWarehouse* se le conoce con el término inglés *drill,* traducido literalmente *drill* significa taladrar**.** Por *drill down* se entiende conseguir datos con un nivel de detalle mayor, profundizar, es la habilidad para poder navegar de lo general a lo particular en la información presentada. Por *roll up* se entiende lo contrario, conseguir datos con un nivel de detalle menor; sintetizar, es agregar un dato según una jerarquía de una dimensión, significa ver menos nivel de detalle, sobre la jerarquía, significa generalizar o sumarizar, es decir, subir en el árbol jerárquico.

## 3.1.4 Carga y mantenimiento del almacén de datos

La carga y mantenimiento de un almacén de datos es uno de los aspectos más delicados y el que más esfuerzo requiere. El sistema encargado del mantenimiento del almacén de datos es el Sistema denominado ETL (*Extraction*, *Transformation*, *Load*)[2] . ETL tiene el cometido de trasladar la información desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, *Datamart* o *DataWarehouse*. La construcción del ETL es responsabilidad del equipo de desarrollo del almacén de datos y se realiza específicamente para cada almacén de datos. Un ETL también se puede cimentar efectuando programas específicos, así como también se puede realizar adaptando herramientas genéricas (por ejemplo *triggers*), herramientas de migración o utilizando herramientas más especificas que van apareciendo cada vez más frecuentemente [2].

Las tareas llevadas a cabo por los procesos ETL pueden verse esquematizadas en la figura 3.3, y son:

- **Extracción.** La extracción es el primer paso para obtener la información que será introducida en el *DataWarehouse*, la mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Para realizarse la extracción

[2] Existen traducciones diversas en castellano, como ETC (Extracción, Transformación, Carga) o ETT (Extracción, Transformación, Transporte).

deben conocerse y comprenderse los orígenes de los datos, y copiar los datos necesarios para procesarlos en las siguientes etapas de carga. Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba [22]. De no ser así los datos son rechazados.

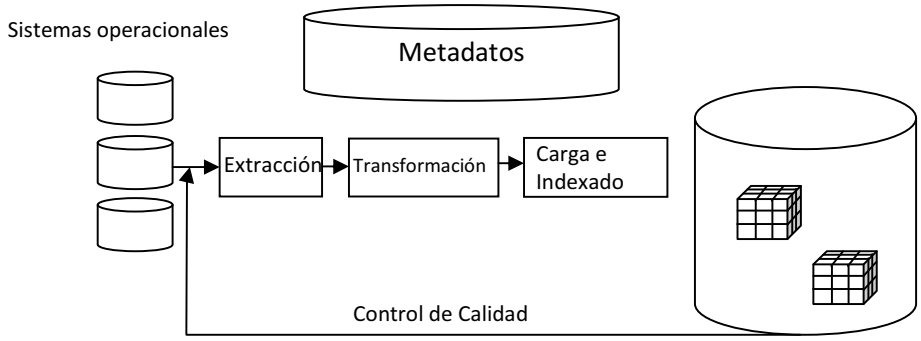



Figura 3.3: Esquema de los procesos ETL.

- **Transformación.** Una vez extraídos los datos existen diferentes tipos de transformaciones posibles sobre ellos:
    - **Corrección de errores en los datos**, por ejemplo, errores tipográficos al introducir los datos, introducción de datos ausentes, y la conversión (*parse)* de los datos para adecuarlos a formatos estándar.
    - **Combinación de fuentes de datos** mediante búsquedas exactas por atributos clave o por búsquedas difusas a partir de atributos que no son claves. Estas búsquedas de información se conocen como *look up.*
    - **Creación de claves**: en general se recomienda crear claves primarias nuevas para todas las tablas que vayan creando en el almacenamiento intermedio o en el almacén de datos [2]. La creación de claves representativas debe garantizar la integridad referencial entre las tablas de hechos y las dimensiones.
    - **Obtención de agregados**: si se sabe que cierto nivel de detalle no es necesario en ningún caso, una primera fase de agregación se puede realizar aquí, para ayudar a acelerar el rendimiento de consultas comunes.
- **Carga e indexado.** Una vez finalizado el proceso de transformación, los datos tienen el formato adecuado para ser introducidos en el *DataWarehouse*. La carga de datos debe

realizarse mediante procesos especiales para grandes volúmenes de datos, mucho más eficientes que las cargas registro a registro. Una vez introducida la información en el *Datamart* correspondiente, deben generarse los índices que permitirán acelerar las consultas sobre el *DataWarehouse*.

- **Control de calidad.** Una vez que se han cargado todos los datos y creados los índices y los agregados en cada *Datamart*, antes de hacer accesible la información a los usuarios debe asegurarse la calidad de la información introducida. Para ello se definen 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.

## 3.1.5 Almacenes de datos y Minería de Datos

Los almacenes de datos pueden utilizarse de muy diferentes maneras, y pueden agilizar muchos procesos diferentes de análisis. En la figura 3.4, se pueden observar las distintas aplicaciones y usos que se puede dar a un almacén de datos: herramientas de consultas e informes, herramientas EIS, herramientas OLAP y herramientas de Minería de Datos.

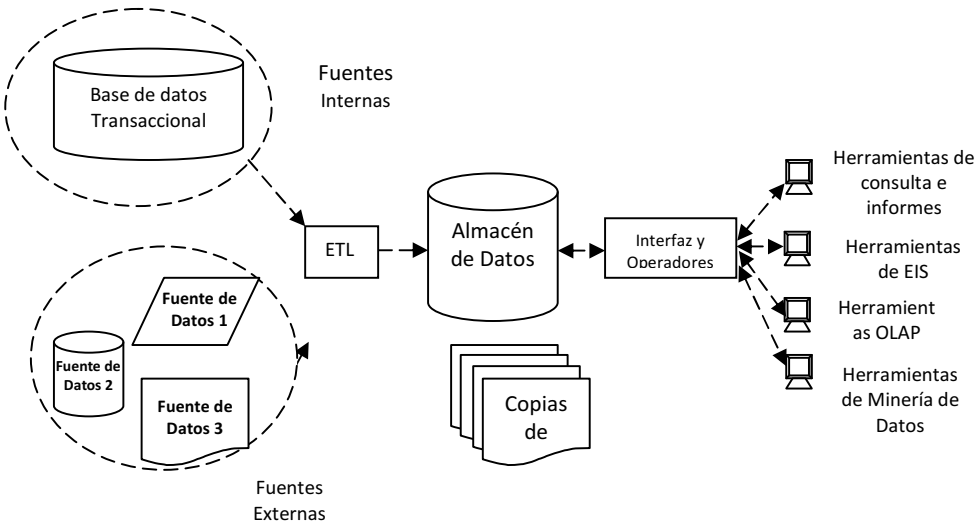



Figura 3.4: Perspectiva general y usos de un almacén de datos [2].

La variedad de usos mostrados en la figura anterior, sugiere la existencia de diferentes grupos de usuarios: analistas, ejecutivos, investigadores, etc. Según el carácter de estos usuarios se les puede catalogar en dos grandes grupos:

- "Picapedreros" (también conocidos como "granjeros"): estos usuarios se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc. [23].
- "Exploradores": encargados de encontrar nuevos patrones significativos utilizando técnicas OLAP o de Minería de Datos.

Los almacenes de datos no son imprescindibles para hacer extracción del conocimiento a partir de datos. Sin embargo, las ventajas de organizar un almacén de datos se amortizan sobradamente a medio y largo plazo. Esto es especialmente evidente cuando las cantidades de datos son exorbitantes o provienen de fuentes heterogéneas o se van a querer combinar de maneras arbitrarias y no predefinidas [2]. En gran medida los almacenes de datos facilitan la limpieza y transformación de datos (en especial al generar "vistas minables" en tiempo real) [2].

## 3.2 Minería de Datos

El sueño del hombre a través de la historia de la computación ha sido desarrollar sistemas inteligentes para el manejo de la información en sistemas de cómputo [24]. La Minería de datos (MD) es una disciplina que combina Técnicas de Inteligencia Artificial, Aprendizaje Computacional, Probabilidad, Estadística, y Bases de Datos para extraer información y conocimientos útiles desde grandes cantidades de datos. Comúnmente el proceso de Minería de Datos convierte datos en conocimiento, en algunos casos se llega a decir que el objetivo es extraer "verdad" a partir de "basura" [26]. Históricamente, a la noción de encontrar patrones útiles en los datos se le ha dado una gran variedad de nombres, como Minería de Datos, descubrimiento de conocimiento en bases de datos (KDD, *Knowledge Discovery in Databases* por sus siglas en inglés), Hallazgo de la Información, Recolección de Información, etcétera. Dichos términos se han ganado la popularidad en las bases de datos En un sentido estricto la Minería de Datos y el KDD no son conceptos equivalentes [27], como se comprenderá en las siguientes secciones.

## 3.2.1 Introducción a la Minería de Datos

La tecnología actual nos permite capturar y almacenar una gran cantidad de datos, para después tratar de encontrar patrones, tendencias y anomalías, lo cual representa uno de los grandes retos de la vida moderna. Algunos de los factores que contribuyen a la generación masiva de datos son: la automatización de procesos (en general), códigos de barras, avances tecnológicos en almacenamientos de información y abaratamiento de los precios de memoria. En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizados de forma manual. El especialista en la materia, analiza los datos y elabora un informe o hipótesis que refleje las tendencias o pautas de los mismos. Esta forma de actuar es lenta, cara y altamente subjetiva. El análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la enorme abundancia de los datos desborda la capacidad humana de comprenderlos sin la ayuda de herramientas potentes. Así el principal cometido de la Minería de Datos es resolver problemas analizando los datos presentes en las bases de datos.

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 (un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados.

En la práctica, los modelos pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos plantean estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que se refiere como variables dependientes 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. Esto se explicará más a detalle en las siguientes secciones, así como también algunas tareas de Minería de Datos que producen modelos predictivos como lo son la clasificación y la regresión, y las que dan lugar a modelos descriptivos como por ejemplo: agrupamiento, las reglas de asociación y el análisis correlacional.

### 3.2.1.1 Tipos de Tareas y Aplicaciones

Algunas de las tareas más destacadas de la Minería de Datos se pueden aplicar en numerosas áreas, prácticamente en todos los movimientos humanos que generan datos:

- Comercio y banca:
    - Segmentación de clientes [28].
    - Previsión de ventas y análisis de riesgo [28].
    - Obtención de patrones de uso fraudulento de tarjetas de crédito.
- Medicina y Farmacia:
    - Diagnóstico de enfermedades.
    - Efectividad de los medicamentos.
    - Detección de pacientes con riesgo de sufrir una patología concreta [2].
    - Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones [2].
- Seguridad y detección de fraude:
    - Reconocimiento facial.
    - Accesos a redes no permitidas.
- Astronomía:
    - Identificación de nuevas estrellas y galaxias [28].
    - Clasificación de cuerpos celestes [2].
- Educación
    - Selección o captación de estudiantes [2].
    - Detección de abandonos y fracaso [2].
    - Estimación del tiempo de estancia en la institución [2].
- Geología, minería, agricultura y pesca:
    - Identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satélites [28].
- Ciencias Ambientales:
    - Identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales.
- Medio Ambiente
    - Predicción de las condiciones del clima, con respecto de la temperatura, precipitación, viento, humedad, etc.
    - Predicción de áreas en incendios forestales.

En todos estos ejemplos se muestra 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 toma de decisiones en dicho entorno.

## 3.2.2 Relación de la Minería de Datos con otras disciplinas

La Minería de Datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello, la investigación y los avances en la Minería de Datos se nutren de los que se producen en estas áreas relacionadas.

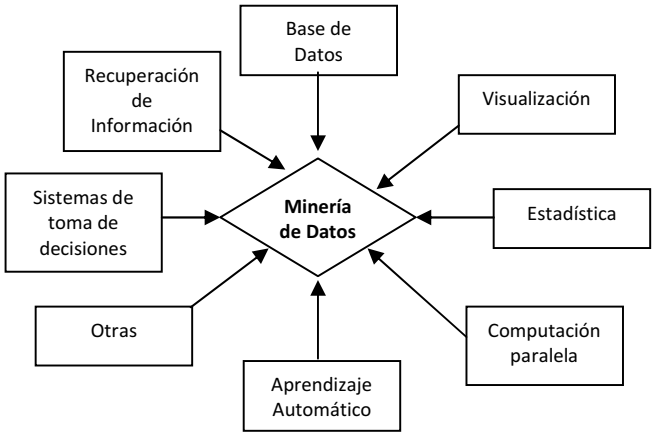



Figura 3.5: Disciplinas que contribuyen a la Minería de Datos [2].

Algunas de las disciplinas más destacadas (Figura 3.5), se describen a continuación:

- Las bases de datos: los almacenes de datos y el procesamiento analítico en línea (OLAP) están relacionados ampliamente con la Minería de Datos, donde el procesamiento OLAP, no se trata únicamente de obtener informes avanzados, como se incluyen en muchas herramientas de *Bussiness Intelligence*, sino que se refieren a la extracción de conocimiento novedoso y comprensible [28].
- La recuperación de información: consiste en la obtención de información relevante desde los datos textuales, donde una tarea típica es encontrar documentos a partir de palabras claves, lo cual puede verse como un proceso de clasificación de los documentos en función de estas palabras clave.

- La estadística: en la Minería de Datos se utilizan mucho los conceptos, algoritmos y técnicas de esta disciplina. Por mencionar algunos tenemos, la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, etcétera.
- El aprendizaje automático: es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender, por medio de algoritmos y programas; constituye junto con la estadística, el corazón del análisis inteligente de los datos.
- Los sistemas para la toma de decisión: (DSS por sus siglas en inglés *Decision Support System*) puede adoptar muchas formas diferentes, y en general es un sistema informático utilizado para servir de apoyo, más que automatizar, el proceso de la toma de decisiones.
- La visualización de los datos: en general son una herramienta para el análisis de los datos. Estas herramientas permiten al usuario descubrir, intuir o entender patrones que serían más difíciles de "ver" a partir de descripciones matemáticas o textuales de los resultados [2].

En la actualidad se puede afirmar que la Minería de Datos ha mostrado la validez de una primera generación de algoritmos mediante diferentes aplicaciones al mundo real [28].

### 3.2.3 La Minería de Datos y la Metodología KDD

La Minería de datos y el Descubrimiento del Conocimiento (KDD) es un campo de rápido crecimiento en la investigación. Su popularidad se debe a la demanda creciente de herramientas que ayudan a revelar y comprender información oculta en enormes cantidades de datos. Los datos se generan a diario por las agencias federales, los bancos, las empresas de seguros, tiendas minoristas y en la WWW. Esta explosión de datos se produce a través del uso creciente de computadoras, escáneres, cámaras digitales, códigos de barras, etc. Actualmente contamos con una rica fuente de datos, almacenados en base de datos, almacenes y otros repositorios de datos, que están disponibles pero no fácilmente analizables. Lo que se necesita es una metodología clara y sencilla para extraer el conocimiento oculto en los datos. Este capítulo explica el modelo de KDD para posibilitar el fortalecimiento del conocimiento permitir la comunicación entre varias herramientas de Minería de datos, base de datos y repositorios de conocimiento. Este capítulo describe los seis pasos del modelo de proceso del KDD y sus componentes tecnológicos.

### 3.2.3.1 Los Seis pasos del KDD y el proceso de Minería de Datos

El objetivo de diseñar un proceso de KDD es, seguir una serie de pasos que pueden ser seguidos por los practicantes cuando ejecutan sus proyectos. Este modelo debe ayudar a planear y así reducir el costo al detallar procedimientos que se realizan en cada uno de los pasos. Los seis pasos del modelo se describen a continuación:

1. **Comprendiendo el dominio del problema.** En este paso se trabaja en estrecha colaboración con expertos del domino para definir el problema y determinar los objetivos del proyecto, se identifican las personas clave, y se aprende sobre soluciones y temas actuales del problema. Consiste en aprender la terminología específica del dominio. Una descripción del problema, incluyendo sus restricciones, para llevarse a cabo.

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

3. **Preparación de los datos.** Este paso es la clave del éxito, ya que el descubrimiento del conocimiento depende de este proceso, por lo general consume aproximadamente la mitad del esfuerzo de todo el proyecto.

4. **Minería de Datos.** Este paso es otro proceso clave, es el proceso de descubrimiento del conocimiento. Las herramientas de Minería de Datos pueden descubrir nueva información, y su aplicación lleva menos tiempo que la preparación de datos. Este paso implica el uso de herramientas de Minería de Datos, y de la selección de otras si es necesario. Las herramientas de Minería de datos incluyen muchos tipos de algoritmos como por ejemplo: métodos bayesianos, computación evolutiva, aprendizaje automático, redes neuronales, *Clustering* y técnicas de preprocesamiento.

Una de las principales dificultades en este paso es que muchas herramientas de uso común no pueden ampliarse para ser aplicadas a un enorme volumen de datos. Las herramientas están caracterizadas por un aumento lineal en el tiempo de ejecución, dentro de una cantidad fija de memoria disponible. La mayor parte de las herramientas de Minería de Datos, no son escalables, pero hay ejemplos de herramientas de Minería que si lo son, como por ejemplo el *Clustering*, el aprendizaje automático y las reglas de asociación.

5. **Evaluación del Conocimiento Descubierto.** Este paso incluye la interpretación de los resultados por parte de los expertos, corresponde a la evaluación si la información es verdaderamente novedosa e interesante, y se lleva un control del impacto del conocimiento descubierto. En todo proceso de KDD se pueden volver a examinar los datos, para identificar alternativas de acción y así mejorar los resultados.

6. **Uso del conocimiento descubierto.** Este paso se encuentra en manos de los propietarios de las bases de datos. Se trata de planificar dónde y cómo el conocimiento descubierto se utilizará. El área de aplicación en el dominio actual debe ampliarse a otros dominios dentro de una organización. Se debe crear un plan para vigilar el conocimiento descubierto y todo el proyecto debe ser documentado.

El grupo de CRISP-DM (*Cross-Industry Standard Process for Data Mining*) tomó la iniciativa al utilizar del proceso de KDD. El proyecto incluyó dos elementos indisociables (con el apoyo de varias empresas de automoción, aeroespacial, telecomunicaciones, consultoría, seguros, almacenamiento de datos, desarrollo de herramientas de Minería de Datos) de cualquier proceso de KDD: herramientas de Base de Datos y Minería. Dos empresas (OHRA y DaimlerChrysler) crearon aplicaciones a gran escala para validar el modelo de KDD. El objetivo del proyecto fue desarrollar un proceso KDD que ayudara a ahorrar costes en los proyectos, reducir el tiempo del proyecto y adoptar marcos alemanes como parte central de la empresa [32]. El proceso resultante es el que se ilustra de forma simplificada en la figura 3.6:

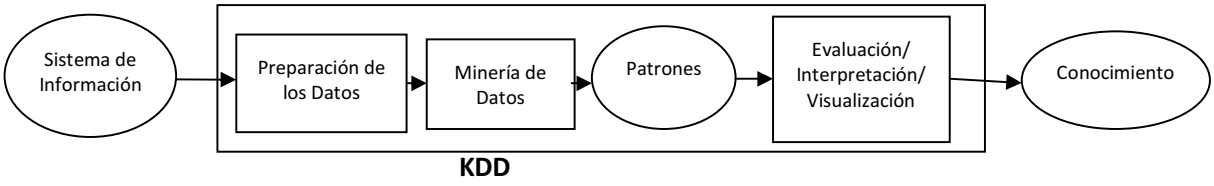



Figura 3.6: Proceso de Extracción de Conocimiento [2].

La definición del proceso anterior clarifica la relación entre el KDD y la MD: el KDD es el proceso global de descubrir conocimiento desde las bases de datos mientras que la MD se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos.

## 3.2.4 Tareas y Técnicas de Minería de Datos

Una tarea de Minería de datos es un tipo de problema de Minería de Datos. Por ejemplo, "clasificar las piezas de una empresa de autopartes en: óptimas, defectuosas, reparables y defectuosas irreparables" es una tarea. Concretamente el tipo de tarea es de clasificación. Está tarea se podría resolver mediante árboles de decisión o redes neuronales, entre otros métodos.

Dentro de las tareas de Minería de Datos existen tipos, cada una de las cuales se puede considerar como un ejemplo de problema a ser resuelto por un algoritmo de Minería de Datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra [2]. Tal y como se ha mencionado en la sección 2.2 las tareas pueden ser predictivas o descriptivas. Entre las tareas predictivas encontramos la clasificación y la regresión, mientras que el agrupamiento (*clustering*), las reglas de asociación secuenciales y las correlaciónales son tareas descriptivas. A continuación se hace una breve descripción de cada una de estas tareas [2].

## 3.2.4.1 Tareas Predictivas

Este tipo de tareas se conoce también como "aprendizaje supervisado", son problemas y técnicas en los que hay que predecir uno 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 se definen varias técnicas predictivas:

- La **clasificación (o discriminación[3])** es quizá la tarea más utilizada. En esta técnica, cada instancia (o registro de la base de datos) pertenece a una clase, la cual se indica mediante el valor de un atributo que se llama clase de instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase. El objetivo es

[3] El término discriminación se utiliza, fundamentalmente en la Estadística

predecir la clase de nuevas instancias de las que se desconoce la clase [2]. Esta técnica responde a preguntas tales como: ¿Cuál es el riesgo de conceder un crédito a este cliente? ¿Dado este nuevo paciente qué estado de la enfermedad indican sus análisis?, entre otras. Existen variantes de la tarea de clasificación, como son el aprendizaje de "*rankings*", el aprendizaje de preferencias, el aprendizaje de estimadores de probabilidad, etc.

En esta técnica los ejemplos se presentan como un conjunto de pares de elementos de dos conjuntos, $\delta = \{< e, s >: e \in E, s \in S\}$ 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, $\delta$ se denomina conjunto de datos etiquetado. El objetivo es aprender una función $\alpha: E \to S$, denominada clasificador, que represente la correspondencia existente en los ejemplos, es decir, para cada valor de $E$ se tiene un único valor para $S$. Además, $S$ es nominal, es decir, puede tomar un conjunto de valores $c_1, c_2, \dots, c_{,m}$, denominados clases (cuando el número de clases es dos, se tiene lo que se llama clasificación binaria). 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 $e$ [2].

- **Clasificación suave**: la presentación del problema es la misma que la de la clasificación, pares de elementos de dos conjuntos, $\delta = \{< e, s >: e \in E, s \in S\}$. Además de la función $\alpha: E \to S$, se aprende otra función $\theta: E \to R$ que significa el grado de certeza de la predicción hecha por la función α. Este tipo de extensión permite realizar otras aplicaciones, como son los *rankings* de predicciones o la selección de los $n$ mejores ejemplos.

- **Estimación de probabilidad o clasificación:** se trata, de una generalización de la clasificación suave. La presentación del problema es la misma que la clasificación normal y suave, pares de elementos de dos conjuntos $\delta = \{< e, s >: e \in E, s \in S\}$. La función a aprender sin embargo, es distinta de la clasificación y la función suave. Se trata de aprender exclusivamente $m$ funciones $\theta_i: E \to R$, donde $m$ es el número de clases. Es decir, cada función a aprender retorna para cada ejemplo $m$ un valor real de $p_i$. Cada uno de estos valores $p_i$ se denomina probabilidad de la clase $i$ [2].

- La **regresión** está es una de las tareas reinas de la Minería de Datos y que ha evolucionado conjuntamente, desde mediados de los 90, con la propia Minería. Esta la técnica es de las más utilizadas para formar relaciones entre datos, también corresponde a las tareas de tipo predictivas. Rápida y eficaz pero insuficiente para espacios multidimensionales donde puedan relacionarse más de dos variables.

El conjunto de evidencias son correspondencias entre dos conjuntos $\delta: E \to S$ donde $S$ es el conjunto de valores de salida. Al igual que en la clasificación, los ejemplos, al ir acompañados de un valor de $S$, se denominan comúnmente ejemplos etiquetados y $\delta$ es un conjunto de datos etiquetado. El objetivo es aprender una función $\delta: E \to S$ que represente la correspondencia existente en los ejemplos, es decir, para cada valor de $E$ se tiene un único valor para $S$. La diferencia con respecto a la clasificación es que $S$ es numérico, es decir, puede ser un valor entero o real [2].

- **Categorización:** esta técnica trata no sólo de aprender una función, sino una correspondencia. Es decir, cada ejemplo de $\delta = \{< e, s >: e \in E, s \in S\}$, así como la correspondencia a aprender $\alpha: E \to S$, pueden asignar varias categorías a un mismo $e$, a diferencia de la clasificación, que sólo asigna una y sólo una. Expresado de otra manera, un ejemplo podría tener varias categorías asociadas. Ejemplos: dado un conjunto de documentos, asignar categorías de los temas que trata cada documento, dados un conjunto de perfiles de clientes, determinar los productos que puedan comprar…. La categorización se puede presentar también en forma de categorización suave (cada categoría asignada va acompañada de su certeza) o en forma de un estimador de probabilidades (se estima una probabilidad para todas las categorías), en este caso la suma de probabilidades puede ser mayor que 1 [2].

- **Preferencias o priorización**: esta técnica trata de determinar a partir de dos o más ejemplos un orden de preferencia. La definición formal es más compleja. Cada ejemplo es en realidad una secuencia $< e_1 e_2, \dots, e_k >, e_i \in E, k \geq 2$ donde el orden de la secuencia representa la predicción. Un conjunto de datos para este problema es, por tanto, un conjunto de secuencias $\delta = \{< e_1 e_2, \dots, e_k >: e_i \in E\}$. Otra manera alternativa de presentar los datos es mediante un orden parcial, donde las secuencias sólo tienen dos elementos $(k = 2)$[2].

## 3.2.4.2 Tareas Descriptivas

Estas técnicas o métodos son también llamados “no supervisados”, utilizados frecuentemente cuando una aplicación no está lo suficientemente preparada y no tiene el potencial necesario para una solución predictiva, para descubrir patrones y tendencias en los datos, que permitan explorar las propiedades de los datos examinados. El objetivo de estas técnicas no es predecir nuevos datos sino describir los existentes y obtener beneficio (científico

o de negocio) de ellas. Algunas de las técnicas más delimitadas son las que se describen a continuación:

- El **agrupamiento o *Clustering*** es un procedimiento de reunión de una serie de vectores según criterios habitualmente de distancia; se trata de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Esta tarea es de tipo descriptiva por excelencia y consiste en obtener grupos (*clusters*[4]) "naturales" a partir de los datos. En esta técnica se habla de grupos y no de clases, esto representa una diferencia fundamental con las técnicas de clasificación, ya que en lugar de analizar los datos etiquetados con una clase, los analiza para generar esta etiqueta. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. De lo anterior se forman grupos tales que los objetos del mismo grupo son muy similares entre sí y al mismo tiempo, son muy diferentes a los objetos de otro grupo. Existen diversas técnicas de agrupamiento. Se dividen en dos grandes categorías:

  - Jerárquicas, que construyen una jerarquía de grupos separándolos iterativamente.
  - De particionamiento, en los que el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía.

- Las **correlaciones** son una tarea de tipo descriptiva que se usa para examinar el grado de similitud de los valores de dos variables numéricas. Una fórmula estándar para medir la correlación lineal es el coeficiente de correlación $r$, el cual es un valor real comprendido entre -1 y 1. Si $r$ es 1 (respectivamente, -1) las variables están perfectamente correlacionadas (perfectamente correlacionadas negativamente), mientras que si es 0 no hay correlación. Esto quiere decir que cuando $r$ es positivo, las variables tienen un comportamiento similar (ambas decrecen o crecen al mismo tiempo) y cuando $r$ es negativo, si una variable crece la otra decrece.

- Las **reglas de asociación** están consideradas como una de las tareas reinas de la Minería de datos y que ha evolucionado con la propia Minería desde mediados de los 90. La técnica tiene como objetivo detectar asociaciones comunes entre elementos (por ejemplo, quien compra cerveza suele comprar también palitos salados). Este tipo de estudios reciben el

[4] Un clúster es la colección de registros que son similares entre sí, y distintos a los registros de otro clúster.

nombre de análisis de asociaciones, o análisis de vínculos (*link analysis*), aunque este término también se utiliza en el agrupamiento jerárquico. [2].

El propósito de la técnica es identificar relaciones no explícitas entre atributos *categóricos*. La formulación más común es del estilo "si el atributo $x$ toma el valor $d$ entonces el atributo $y$ toma el valor $b$". Las reglas de asociación tienen usos típicos en:

- Análisis de la cesta de compra
- Identificación de productos que son comprados juntos y es como se logra ofrecer recomendaciones al comprador: ¿has comprado cerveza, seguro que no quieres palitos salados?
- Información que puede utilizarse para ajustar los inventarios, y así tener una buena organización física del almacén.
- En campañas publicitarias.

Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados. Un aspecto importante de las reglas, es que estas se evalúan usando dos parámetros, precisión y soporte (cobertura).

Un caso especial de reglas de asociación recibe el nombre de **reglas de asociación secuenciales**, se usa para determinar patrones secuenciales en los datos. Estos patrones se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre los datos se basan en el tiempo. Por ejemplo: en una tienda de venta de electrodomésticos y equipos de audio se analizan las ventas que ha efectuado usando, un análisis secuencial y se descubre que el 30 por ciento de los clientes que compraron un televisor hace seis meses compraron un DVD en los siguientes dos meses.

- **Dependencias Funcionales:** el descubrimiento de dependencias funcionales generalmente se incluye en las tareas con el nombre de "reglas de asociación". En realidad las dependencias funcionales consideran todos los posibles valores (a diferencia de las asociaciones o dependencias de valor). Se definen de la siguiente manera: "dados los valores de $A_i, A_j, \dots, A_k$ se puede determinar el valor de $A_r$". Es decir, el valor de $A_r$ depende o es función de los valores de ciertos atributos $A_i, A_j, \dots, A_k$ [2].

- **Detección de valores e instancias anómalas:** la detección de valores anómalos o atípicos (*outlier detection*) es muy útil para detectar comportamientos anómalos, que pueden sugerir

fraudes, fallos, intrusos o comportamientos diferenciados. La definición de instancia anómala es más general en el sentido de que no sólo se considera un único atributo, sino que se consideran todos. El objetivo de la tarea es encontrar aquellas instancias que no son similares a ninguna (o muy pocas) de las otras instancias. Para esta tarea se utilizan los agrupadores suaves o los estimadores de probabilidad de agrupamiento, ya que si un ejemplo tiene baja probabilidad de agrupamiento con todos los grupos se puede considerar un caso “aislado” y, por tanto, anómalo [2].

- **Arboles de decisión:** es un modelo de predicción utilizado en el ámbito de la Inteligencia Artificial. Dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Son útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclen datos categóricos y numéricos. Esta técnica se usa en tareas de clasificación, agrupamiento y regresión. Los árboles de decisión usados para predecir variables categóricas reciben el nombre de árboles de clasificación, ya que distribuyen las instancias en clases. Los árboles de decisión usados para predecir variables continuas se llaman árboles de regresión.

### 3.2.5 Métodos. Correspondencia entre tareas y métodos

Las tareas anteriormente descritas, requieren métodos, técnicas o algoritmos para resolverlas. Una tarea puede tener muchos métodos para resolverla, se da el caso de que un mismo método (o el mismo tipo de técnica) puede tener un gran abanico de tareas. En esta sección se explican brevemente los tipos de técnicas existentes para llevar a cabo las tareas anteriores. La relación que se exhibe ofrece una perspectiva general de la diversidad de técnicas existentes.

- Técnicas algebraicas y estadísticas: generalmente se basan en expresar modelos y patrones mediante fórmulas 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, de aquí el nombre de técnicas paramétricas. 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.

- 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 para ello el Teorema de Bayes. Algunos algoritmos populares son el clasificador de Naïve Bayes, los métodos basados en Máxima Verisimilitud y el algoritmo EM.

- 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 sólo en aquellos casos en que las frecuencias conjuntas superen cierto umbral.

- Técnicas basadas en árboles 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 ID3/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 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 dan 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.

- Técnicas basadas en núcleo y máquinas 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 muchas variantes, dependiendo del núcleo utilizado y de la manera de trabajar con el margen.

- Técnicas basadas en casos, en densidad o distancia: son métodos que se basan en distancias al resto de los elementos, ya sea directamente, como los vecinos más próximos (los casos similares), de una manera más sofisticada, mediante la estimación de funciones de densidad.

A continuación se muestran de manera ilustrativa, algunas tareas (clasificación, regresión, agrupamiento, reglas de asociación, correlaciones/factorizaciones) y algunas técnicas o algoritmos.

Tabla 3.2. Correspondencia entre tareas y técnicas de la Minería de Datos [2].

| **Nombre** | **PREDICTIVO** | | **DESCRIPTIVO** | | |
|---|---|---|---|---|---|
| | **Clasificación** | **Regresión** | **Agrupamiento** | **Reglas de Asociación** | **Correlaciones y Factorizaciones** |
| Redes Neuronales | X | X | X | | |
| Árboles de decisión: ID3, C4.5, C5.0 | X | | | | |
| Árboles de decisión CART | X | X | | | |
| Otros árboles de decisión | X | X | X | X | |
| Redes de Kohonen | | | X | | |
| Regresión lineal y logarítmica | | X | | | X |
| Regresión logística | X | | | X | |
| Kmeans | | | X | | |
| A priori | | | | X | |
| Naïve Bayes | X | | | | |
| Vecinos más próximos | X | X | X | | |
| Análisis factorial y de comp. Ppales. | | | | | X |
| Twostep, Cobweb | | | X | | |
| Algoritmos genéticos y evolutivos | X | X | X | X | X |
| Máquinas de vectores soporte | X | X | X | | |
| CN2 rules (cobertura) | X | | | X | |
| Análisis discriminante multivariante | X | | | | |

## 3.2.6 Técnicas de Minería de Datos Utilizadas

### 3.2.6.1 *K-Means*

El nombre de K-medias (del inglés *K-Means*) proviene de la representación de cada uno de los *clusters* por la media (o media ponderada) de sus puntos, es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. Cada clúster por tanto es caracterizado por su centro o centroide (Figura 3.7).

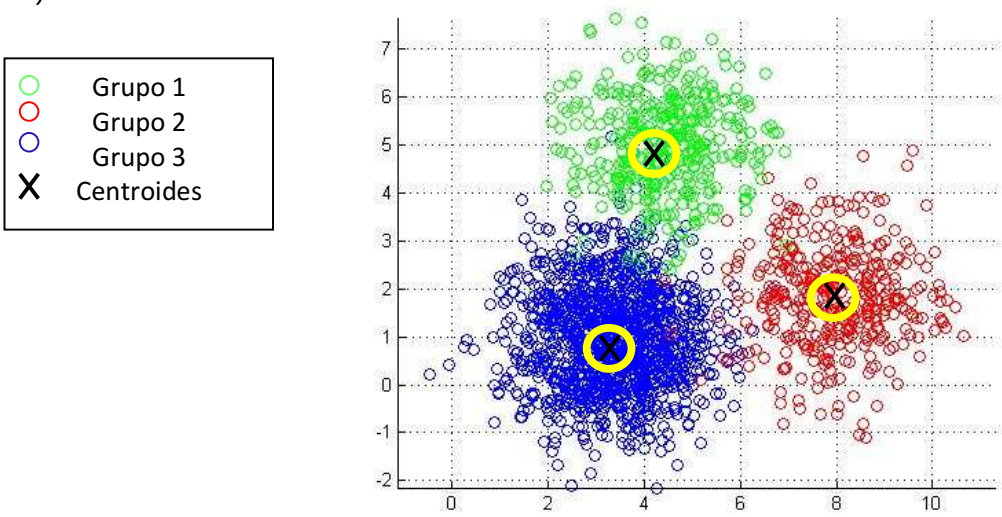



Figura 3.7: Clústeres caracterizados por su centroide.

El algoritmo *K-Means* se trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar, sin etiquetar. Es el más popular entre los métodos de agrupamiento denominados "por partición". La idea del *K-Means* es situar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares [31].

Todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con éstos y asociado a aquél que sea el más próximo, en los términos de una distancia previamente elegida. Normalmente se usa la distancia Euclideana [2].

Las regiones se definen minimizando la suma de las distancias cuadráticas entre cada vector de entrada y el centro de su correspondiente clase, representado por el prototipo correspondiente. El algoritmo puede seguir dos enfoques distintos: KI-medias por lotes (*batch*) y K-medias en línea (*on-line*). El primero se aplica cuando todos los datos de entrada están disponibles desde un principio, mientras que el segundo se aplica cuando no se dispone de todos los datos desde el primer momento, sino que pueden añadirse ejemplos adicionales mas

tarde. Cuando se aplica por lotes, se debe seleccionar arbitrariamente una partición inicial de forma que cada clase disponga de, al menos, un ejemplo. Como la totalidad de los datos están disponibles, los centros de cada partición se calculan con la media de los ejemplos pertenecientes a esa clase. A medida que el algoritmo se va ejecutando, algunos ejemplos cambian de una clase a otra debiendo recalcularse los centros en cada paso, o sea, desplazar convenientemente los prototipos.

El método tiene una fase de entrenamiento, que puede ser lenta, dependiendo del número de puntos a clasificar y la dimensión del problema. Una vez terminado el entrenamiento, la clasificación de nuevos datos es rápida, gracias a que la comparación de distancias se realiza con los prototipos.

El procedimiento es el siguiente:

- Se calcula, para cada ejemplo $x_k$ el prototipo más próximo $A_g$ y se incluye en la lista de ejemplos de dicho prototipo.

$$A_g = \arg min_{A_i}\{d(x_k, A_i)\} \quad \forall i = 1 \dots n$$

- Después de haber introducido todos los ejemplos, cada prototipo $A_k$ tendrá un conjunto de ejemplos a los que representa:

$$l(A_k) = \{x_{k_1}, x_{k_2}, \dots, x_{k_m}\}$$

- Se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos.

$$A_k = \frac{\sum_{i=1}^{m} x_{k_i}}{m}$$

- Se repite el procedimiento hasta que ya no se desplazan los prototipos.

Mediante este algoritmo el espacio de ejemplos de entrada se divide en $k$ clases o regiones, y el prototipo de cada clase estará en el centro de la misma. Dichos centros se determinan con el objetivo de minimizar las distancias cuadráticas euclídeas entre los patrones de entrada y el centro más cercano, es decir minimizando el valor J:

$$J = \sum_{I=1}^{K} \sum_{n=1}^{m} M_{i,n} d_{EUCL}\ (x_n - A_i)^2$$

Donde *m* es el conjunto de patrones, es la distancia euclídea, es el ejemplo de entrada *n,* es el prototipo de la clase *i*, y es la función de pertenencia del ejemplo *n* a la región *i* de forma que vale 1 si el prototipo es el más cercano al ejemplo y 0 en caso contrario es decir:

1 si

0 en caso contrario

## 3.2.6.2 Regresión Lineal

La **regresión lineal**, es un método simple pero frecuentemente utilizado para la tarea de regresión (Figura 3.8). En general, la fórmula para una regresión lineal es

, donde son los atributos predictores e la salida (la variable dependiente). Si los atributos son modificados en la función de regresión por alguna otra función (cuadrados, inversa, logaritmos, combinaciones de variables…), es decir , la regresión se dice no lineal. Se pueden incorporar variantes locales o trasformaciones en las variables predictorias y en la salida, permitiendo flexibilizar este tipo de técnicas. El abanico de técnicas se dispara aún más cuando consideramos técnicas no paramétricas.

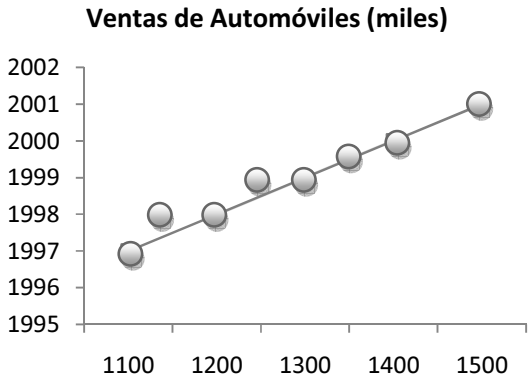



Figura 3.8: Ejemplo de regresión lineal.

## 3.2.6.3 Árboles de decisión

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, etc.

Una de las ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

Un árbol de decisión lleva a cabo un *test* a medida que éste se recorre hacia las hojas para alcanzar una decisión. El árbol de decisión suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un *test* sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, este tipo de nodos es redondo, los demás son cuadrados. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada. Los arboles de decisión poseen:

- Ramas: se representan con líneas.
- Nodos de decisión: de ellos salen las ramas de decisión y se representan por un cuadrado.
- Nodos de incertidumbre: de ellos salen las ramas de los eventos y se representan con un círculo.

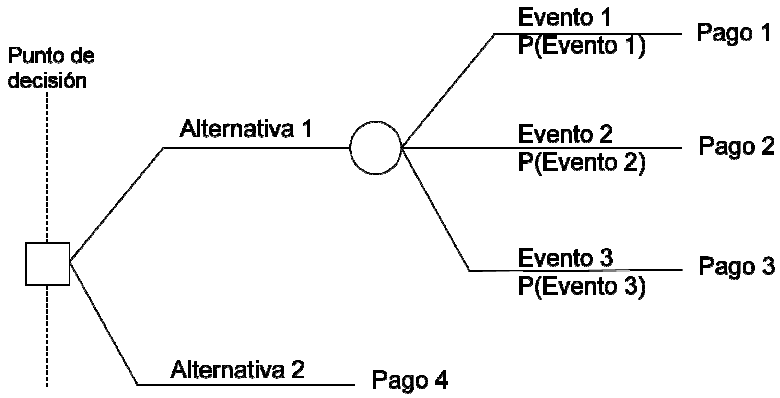



Figura 3.9: Ejemplo de Componentes y Estructura de un árbol de decisión.

Los árboles de decisión tienen un riesgo, el cual se refiere a la variación en los resultados posibles. Mientras más varíen los resultados entonces se dice que el riesgo es mayor. Existen diferentes maneras de cuantificar el riesgo, y una de ellas es la varianza. La varianza se calcula como:

$$\mathrm{var}(X) = \sum_{j=1}^{m} p(X_j) \cdot \left[ X_j - E(X) \right]^2$$

Los árboles de decisión se fundamentan en métodos y algoritmos de aprendizaje, como por ejemplo los sistemas de reglas que son una generalización de los árboles de decisión en el que no se exige exclusión ni exhaustividad en las condiciones de las reglas. La representación en forma de reglas suele ser, en general, más sucinta que la de los árboles, ya que permite englobar condiciones y permite el uso de reglas por defecto.

## 3.2.7 Entorno de Minería de Datos WEKA

WEKA (*Waikato Environment for Knowledge Analysis*) es una herramienta visual de libre distribución (licencia GNU) desarrollada por un equipo de investigadores de la universidad de Waikato (Nueva Zelanda). Como entorno de Minería de datos conviene destacar [2]:

- Acceso a datos: los datos son cargados desde un archivo en formato ARFF (archivo plano organizado en filas y columnas), El usuario puede observar en los diferentes componentes gráficos, información de interés sobre el conjunto de muestras (talla del conjunto, número de atributos, tipo de datos, medias y varianzas de los atributos numéricos, distribución de frecuencias en los atributos nominales, etc.)
- Preprocesado de datos (destacar la gran cantidad de filtros disponibles):
  - Selección de atributos.
  - Discretización.
  - Tratamiento de valores desconocidos.
  - Transformación de atributos numéricos.
- Modelos de aprendizaje:
  - Árboles de decisión (J4.8, versión propia del método C4.5).
  - Tablas de decisión.
  - Vecinos más próximos.

- Máquinas de vectores de soporte (método *sequential minimal optimization*).
- Reglas de asociación (método Apriori).
- Métodos de agrupamiento (K-medias, EM y Cobweb).
- Modelos combinados (*bagging, boosting, stacking,* etc).

- Visualización (la interfaz gráfica se compone de diversos entornos):
  - El entorno *Explorer* permite controlar todas las operaciones anteriores (filtrado, selección y especificación del modelo, diseño de experimentos, etc.).
  - El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. (También es posible acceder directamente a los métodos que implementan dichas tareas e incorporarlos en el código fuente de la aplicación de Minería de Datos que se esté programando.)
  - 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 (carga de datos, preproceso, obtención de modelos, comprobación y visualización de resultados).

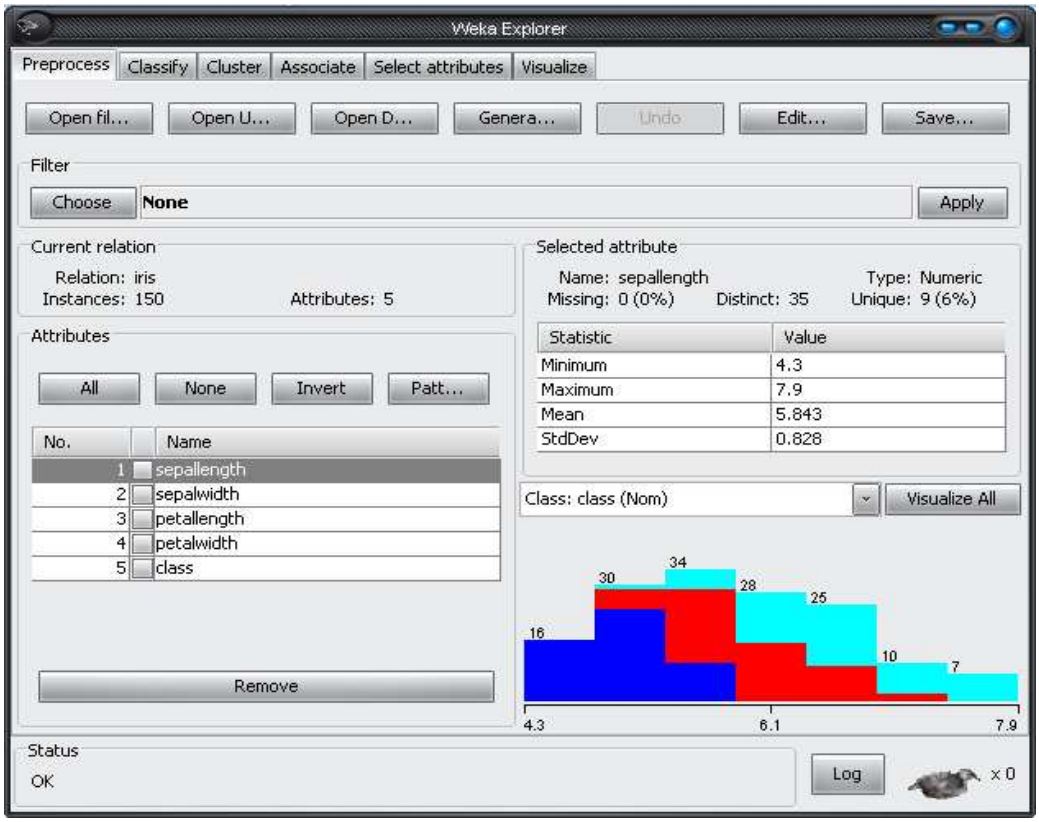



Figura 3.10: Detalle del entorno *Explorer* de WEKA.

### 3.2.8 Conclusiones

El marco Teórico utilizado en esta Tesis, ha ampliado los conocimientos científicos acerca de los *DataWarehouse* y la Minería de Datos, ayudando en la ampliación del problema descrito en la sección 1.1.1. Así como también se efectúo una integración de la teoría existente sobre los conocimientos científicos con investigación y sus relaciones mutuas.

# Capítulo 4

# Análisis y Diseño

El objeto del presente capitulo es mostrar el diseño, construcción e implantación del almacén de datos para el Sistema de Toma de Decisiones acerca del Cambio Climático, posteriormente se describe el procedimiento correspondiente a la aplicación de Técnicas de Minería de datos, en concordancia con lo teóricamente expuesto en el capítulo 3.

## 4.1 Planteamiento y requerimientos

Tomando en cuenta los problemas identificados en la sección 1.1.1 y los objetivos de la investigación planteados en la sección 1.1.2, los requerimientos que se necesitan siguen la metodología del proceso de KDD expuesto en la sección 3.2.3.1**,** como a continuación se muestra:

1. **Preparación de los datos**

    - Integrar y recopilar los datos para determinar las fuentes de información que pueden ser útiles, para la construcción del *DataWarehouse*.
    - Efectuar un proceso de selección, limpieza y transformación de la información, y así eliminar y corregir los datos incorrectos, con esto se decidirá la estrategia a seguir con los datos incompletos. En esta parte se proyectarán los datos y se consideraran únicamente aquellas variables o atributos que son relevantes, con el objetivo de facilitar la tarea propia de la Minería y para que los resultados de la misma sean útiles.

2. **Aplicación de las técnicas de Minería de Datos**

    - En esta fase se decidirá cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elegirá el método a utilizar.

3. **Fase de Evaluación e Interpretación**

- Se evaluarán los patrones y se analizarán por los expertos, y si es necesario se realizarán las fases anteriores para una nueva iteración.

4. **Fase de Difusión y uso**

- Se hará uso del nuevo conocimiento y se hará partícipe de él a todos los posibles usuarios.

A manera de ilustrar los requerimientos anteriores éstos se muestran en la figura 4.1:

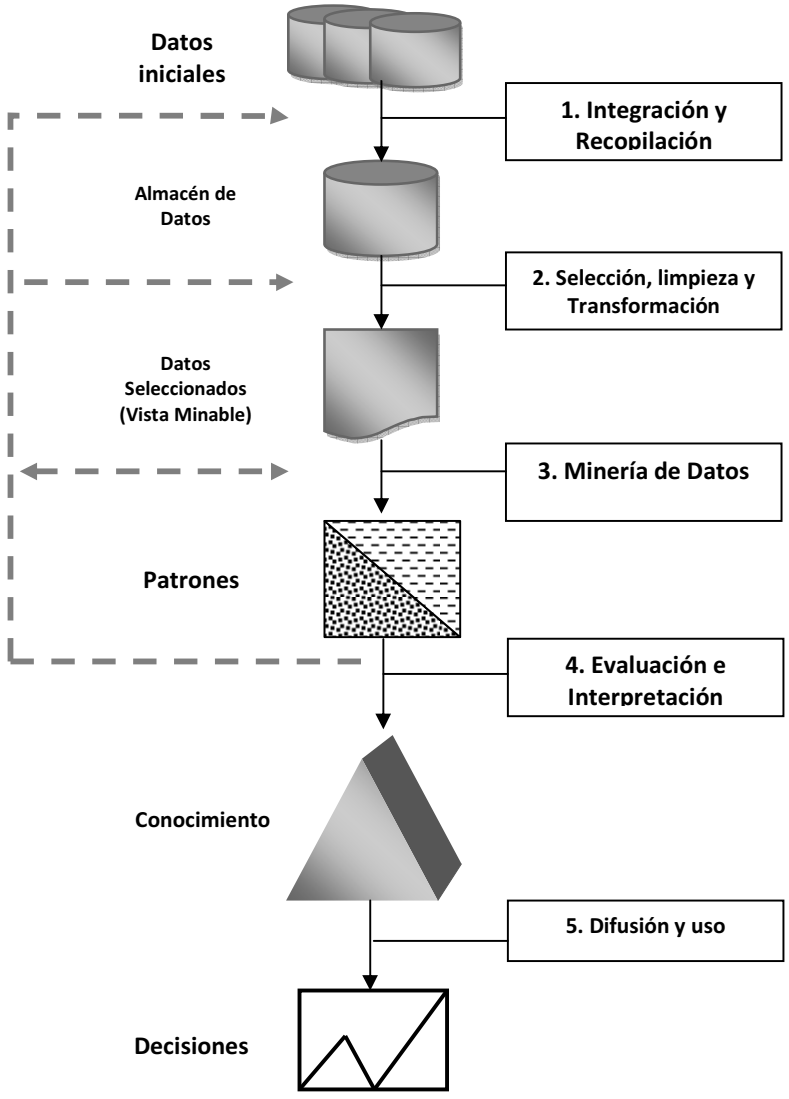



Figura 4.1: Fases del proceso de descubrimiento en bases de datos, KDD.

## 4.2 Fase de Integración y recopilación

En esta sección se describe la integración de los datos suministrados por la SEMARNAT. En junio de 2000 la SEDURBECOP, instaló e inició operaciones de una red de calidad del aire con cuatro estaciones de monitoreo automático, que incluyen la medición de contaminantes como el Ozono, Dióxido de Azufre, PM10[5], y medidas relacionadas con Meteorología (humedad relativa, radiación UV-A, radiación UV-B, velocidad del viento, dirección del viento y presión barométrica). Actualmente, el Sistema Estatal de Monitoreo Ambiental, SEMA, se encuentra a cargo de la SEMARNAT Delegación Puebla.

Los resultados obtenidos de esta integración de medidas de la calidad del aire se muestran en las tablas 4.1 y 4.2, así como también los parámetros y estaciones del Sistema Estatal de Monitoreo Atmosférico de Puebla. Igualmente en la figura 4.2 se presenta un mapa correspondiente a la ubicación de cada una de las estaciones, dentro de Puebla Capital.

**Tabla 4.1** Estaciones del Sistema Estatal de Monitoreo Atmosférico de Puebla

| Estación | Clave |
|---|---|
| Agua Santa | AGS |
| Las Ninfas | NIN |
| Hermanos Serdán | HES |
| Tecnológico | TEC |

**Tabla 4.2** Parámetros del Sistema Estatal de Monitoreo Atmosférico de Puebla

| Alias | Nombre Atributo |
|---|---|
| WS | Velocidad del viento |
| WD | Dirección del viento |
| TEMP | Temperatura |
| HR | Humedad Relativa |
| BPR | Presión |
| UV-A | Radiación UV-A |
| UV-B | Radiación UV-B |
| OZONO | Ozono |
| SO2 | Bióxido de Azufre |
| PM10 | Partículas menores a 10 micras |
| No Estación | Número de Estación |
| Fecha | Fecha : Año, mes, día |
| Hora | Hora |

[5] Las partículas $PM_{10}$ abarcan un amplio espectro de sustancias orgánicas e inorgánicas que se caracterizan por tener un diámetro inferior a 10 micras. Las partículas $PM_{10}$ tienen en el tráfico una de sus principales fuentes, aunque también contribuyen las actividades industriales y algunas situaciones naturales. Las partículas $PM_{10}$ penetran en el tracto respiratorio y generan diversas afecciones respiratorias, agravamiento de cuadros alérgicos y problemas cardiovasculares, además de estar relacionada con el cáncer de pulmón [5].

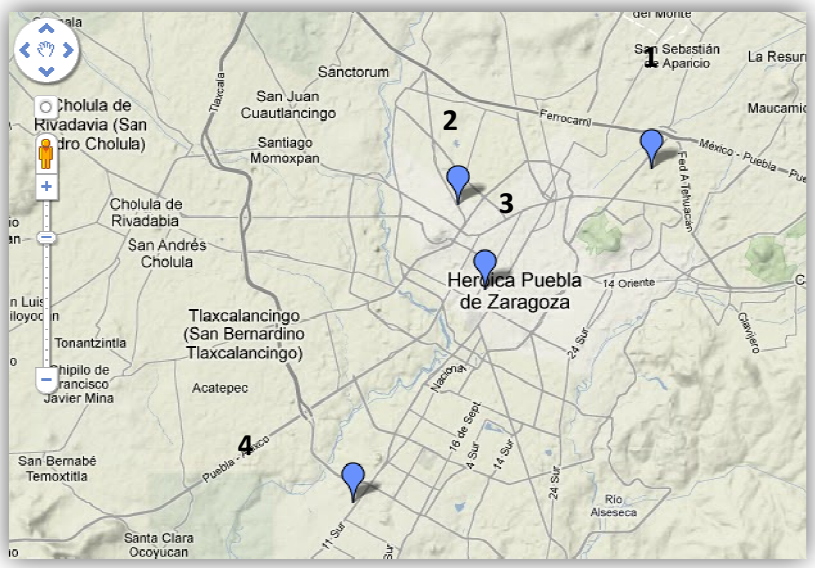



Figura 4.2: Ubicación de estaciones del Sistema Estatal de Monitoreo Atmosférico de Puebla [1].

A continuación se describe la ubicación específica de cada una de estas estaciones de monitoreo de acuerdo a la Figura 4.2:

1. El número “1” corresponde a la estación de monitoreo Tecnológico (TEC), con ubicación en Paseo Morelos No. 940 Col. Esteban Cantú, Puebla, Puebla. CP. 21440 con coordenadas; latitud: 19°4'18.85" y longitud: 98°10'11.38" [1].
2. El número “2” corresponde a la estación de monitoreo Hermanos Serdán (HES), con ubicación en Boulevard Hermanos Serdán y Boulevard San Felipe Hueyotlipán, Col. Aquiles Serdán, Puebla, Puebla. C.P. 72029 con coordenadas; latitud: 19° 3'45.89", y longitud: 98°13'17.30" [1].
3. El número “3” corresponde a la estación de monitoreo Las Ninfas (NIN), con ubicación en 23 Poniente y 15 Sur, Col. Santiago, Puebla, Puebla. C.P. 72270 con coordenadas; latitud: 19°2'28.75" y longitud: 98°12'51.08" [1].
4. El número “4” corresponde a la estación de monitoreo Agua Santa (AGS), con ubicación en Prolongación 11 Sur, Col. Agua Santa, Puebla, Puebla. C.P. 72000 con coordenadas; latitud: 98°14'58.45" [1].

Los datos suministrados por la SEMARNAT están conformados por 210348 registros provenientes de las 4 estaciones de Puebla. Cada registro contiene 10 tipos de mediciones.

## 4.3 Fase de Selección, limpieza y transformación

La calidad del conocimiento descubierto no sólo depende del algoritmo de Minería utilizado, sino también de la calidad de los datos minados. Por ello, después de la recopilación, el siguiente paso en el proceso de KDD es seleccionar y preparar el subconjunto de datos que se va a minar, los cuales constituyen lo que se conoce como vista Minable.

En este proceso se eliminaron registros de 5 o más mediciones perdidas o faltantes, también se eliminaron valores *outliers* (valores que no se ajustaron al comportamiento general de los datos). Este procedimiento se realizó utilizando una visualización previa de histogramas conseguidos con la herramienta Weka (ver figura 4.3). Para llenar completar registros en algunas mediciones, se realizo un cálculo de promedio. Finalmente se obtuvieron 191 435 registros completos. Estos registros fueron aprobados por los expertos del clima en la SEMARNAT.

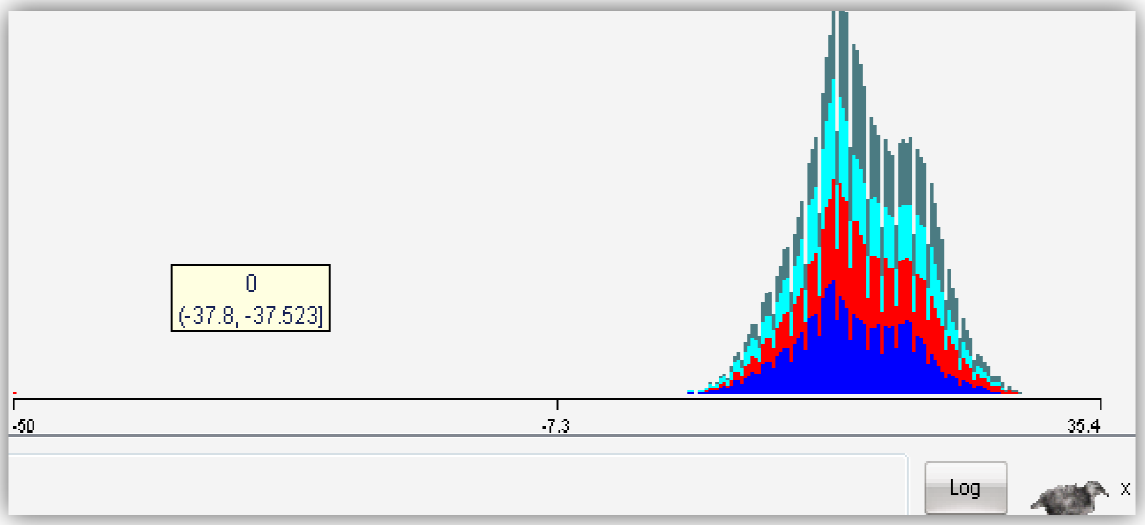



Figura 4.3: Visualización de datos *outliers* con Weka.

En la figura 4.3, se muestra el histograma del parámetro de temperatura, donde se visualiza la existencia de datos de hasta -50 grados centígrados, este y otros datos fueron eliminados del conjunto.

Una vez seleccionados los atributos con respecto de la tabla anterior, se procedió a realizar una preparación de los datos, es decir, la construcción automática de nuevos atributos, con objeto de que estos nuevos atributos hagan más fácil el proceso de Minería. Por ejemplo los atributos de fecha y hora fueron discretizados (se pueden tratar como atributos categóricos) con un número más pequeño de valores (Figura 4.4).

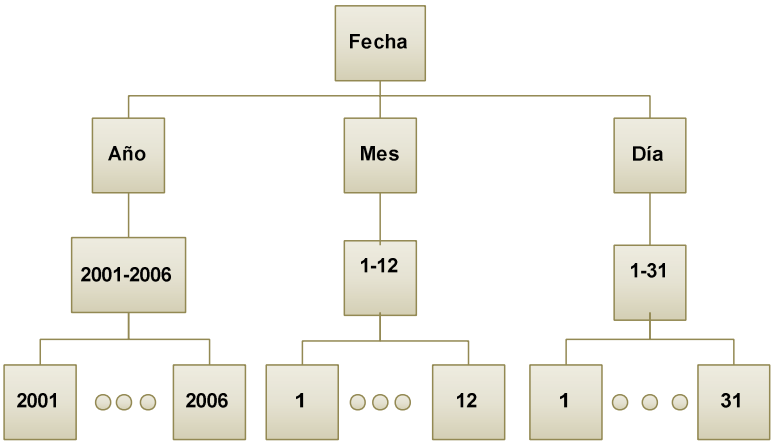



Figura 4.4: Ejemplo de discretización del atributo Fecha.

Como se muestra en la figura 4.4, la discretización para el atributo Fecha, se ha discretizado en tres intervalos a los que se les ha asignado los valores discretos de año, mes y día.

## 4.4 Construcción del Almacén de Datos

Una vez realizado el proceso de selección, limpieza y transformación, se procedió la construcción del *DataWarehouse* utilizando la herramienta SQL *Server Business Intelligence Development Studio.* A continuación se describe el procedimiento de construcción del cubo de datos.

Como primer paso de este proceso se incorporó el conjunto de datos al SQL *Server,* mejor conocido como Analizador de Servicios (*Analysis Services*), este permite visualizar la estructura del origen de datos mediante el uso de vistas. Una vez que lo anterior es procesado y visualizado, se puede proceder con la creación de las dimensiones y medidas del cubo, acciones que permiten agilizar, consultar y modificar la información para una posterior toma de decisiones.

A partir de una base de datos, ESTACIONES-SERMARNAT (la cual fue diseñada en SQL-Server 2008), se generó un diagrama E-R, el cual se compone de las siguientes tablas: 1. Estación, 2. Tiempo y 3. Mediciones (Figura 4.5).

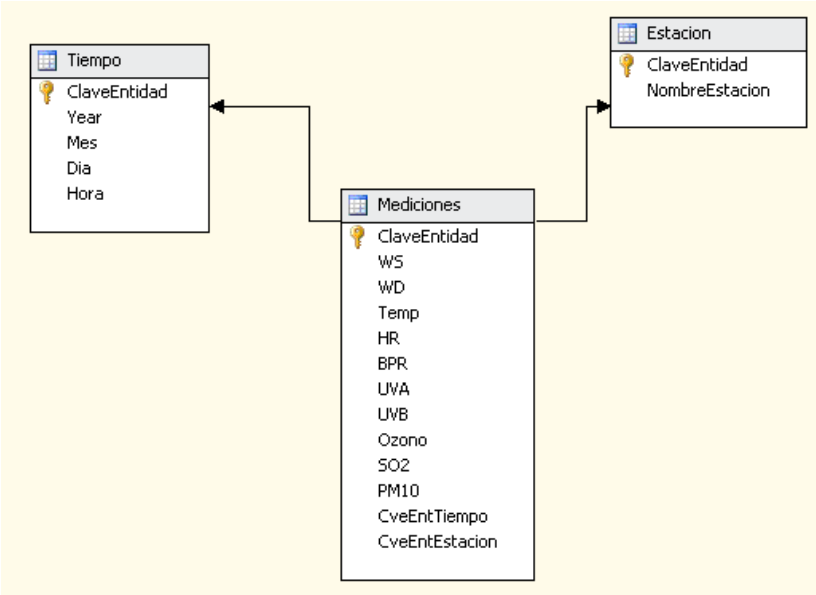



Figura 4.5: Modelo E-R de la base de datos ESTACIONES-SEMARNAT.

Se puede observar en la figura 4.5, que existe lo que se denomina una tabla de hechos o tabla principal, con el nombre de Mediciones, encargada de recoger todas las características o atributos que van a permitir relacionarse con las demás entidades.

La elaboración y construcción de cubos en SQL Server 2008, comienza por la selección de un origen de datos, donde y mediante selección de atributos y previo establecimiento de jerarquías de las base de datos, se procede de forma automática con la creación de los respectivos cubos.

A continuación se describen cada una de las tablas de la figura anterior.

- Medidas: son el producto final que se obtiene una vez seleccionadas las variables de dimensión contra las variables de hechos, las medidas incluidas en el cubo son:
  - Dimensiones: son el resultado de haber seleccionado unas variables de tipo dimensión y algunas de las medidas disponibles que van a conformar el cubo final.

El resultado final, se puede apreciar en la figura 4.6, la cual muestra las dimensiones finales del cubo.

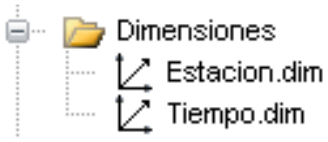



Figura 4.6: Dimensiones finales del cubo.

En la figura anterior, se puede apreciar que las dimensiones generadas, se identifican con tres ejes de análisis a diferencia de los hechos que sólo usan dos ejes.

La vista es la representación gráfica de las tablas y sus respectivas relaciones. Las vistas obtienen información de las tablas, las cuales fueron creadas en los orígenes de los datos. En la figura 4.7, se muestran de manera ilustrativa las tablas del almacén de datos, denominado Contaminacion.cube.

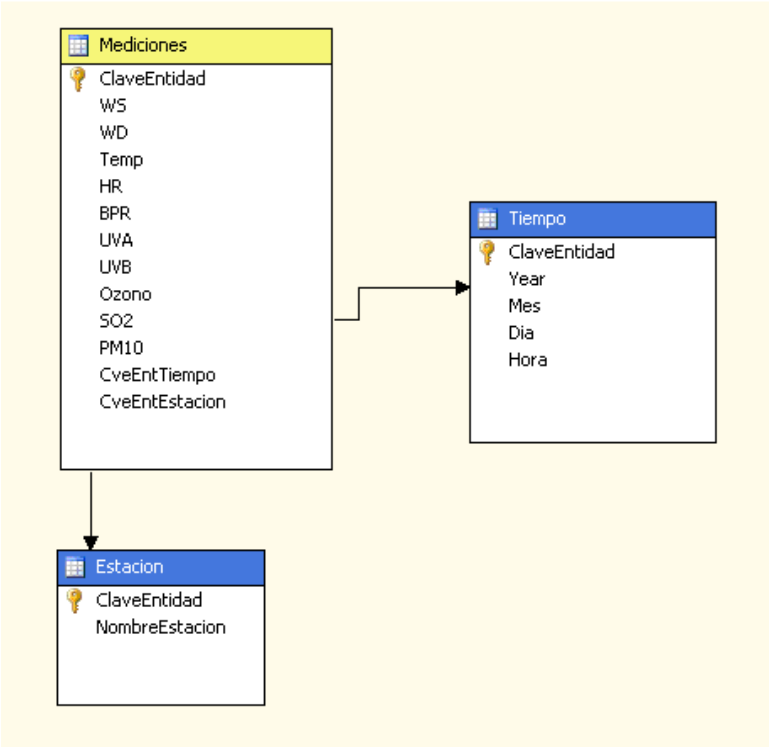



Figura 4.7: Vistas y Tablas.

## 4.5 Fase de Minería de Datos

El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas.

A continuación se describen las decisiones tomadas, para minar el conjunto de datos conseguido en el paso anterior:

### 1. Elección de tarea de Minería de Datos.

A continuación se describen brevemente algunas técnicas aplicadas al conjunto de datos:

- ***K-Means*:** fue la primera técnica aplicada, (a partir de un número k de *clusters*, obtenidos por medio de una previa visualización de los datos, que sugiere un número k=3 *clusters* (Figura 4.8)). Dado que el conjunto de datos es numérico, la aplicación de *K-Means* encaja perfectamente. La técnica también permite dividir los datos en grupos teniendo en cuenta el criterio de la distancia Euclideana.

- **Regresión Lineal**: esta técnica fue aplicada en segundo lugar, con el fin de predecir el comportamiento de la Capa de Ozono en relación con las condiciones atmosféricas. La técnica es capaz de generar un modelo para explicar el comportamiento de la Capa de Ozono, en determinadas condiciones. Además, la estructura del conjunto de datos, permite la aplicación de esta técnica.

- **Arboles de decisión**: finalmente se aplicó la técnica de árboles de decisión, en particular los de tipo J48. Los árboles se utilizaron (tomando en cuenta los resultados de la regresión lineal), para responder una pregunta: ¿Se puede determinar los niveles de Ozono, con relación a las variables cubiertas por la regresión lineal?

  Las técnicas anteriores, fueron aplicadas, utilizando la herramienta WEKA, la cual es comúnmente usada para minar sobre conjuntos de datos.

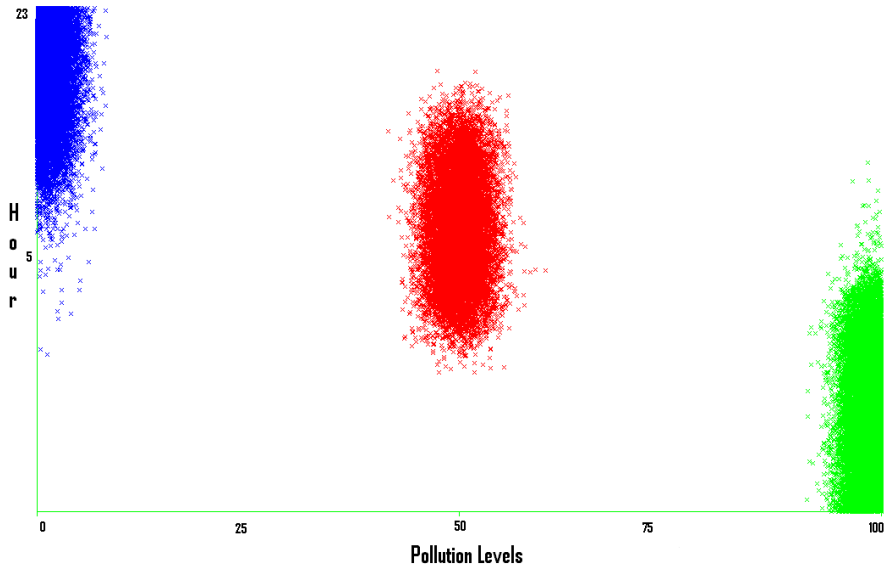



Figura 4.8: Grupos de contaminación durante el día.

## 4.6 Conclusiones

En este capítulo se presentó el diseño, construcción e implantación del almacén de datos para el sistema de Toma de decisiones acerca del Cambio Climático, siguiendo los pasos de la metodología KDD. Así como también se describieron las Técnicas de Minería de Datos utilizadas para la explotación del almacén de datos.

# Capítulo 5

# Resultados

Esta sección muestra los resultados de la ejecución de cada Técnica de Minería, para el conjunto de datos descrito en la sección anterior 4.2.

## 5.1 *Clustering*

Como primer experimento, se aplicó la técnica de *K-Means* al conjunto de datos, usando un número k=3 de grupos (obtenidos por previa visualización). En la vista preliminar, se observó que existen factores que discriminan a los grupos, tales como la radiación UV-A y UV-B, dichas mediciones sólo se producen durante el día, donde la temperatura es alta y ligeramente inferior en la noche. Una descripción del comportamiento de los contaminantes y las mediciones meteorológicas a lo largo del día y para cada estación, se presentan en las tablas 5.1 a 5.7. Las columnas corresponden a las 4, 13 y 20 horas, donde los parámetros de las mediciones fueron significativos.

Tabla 5.1. Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 1.

| **Hora** | **4** | **13** | **20** |
|---|---|---|---|
| WD | 171.9229 | 207.9924 | 175.9858 |
| WS | 4.2333 | 5.8775 | 5.6153 |
| Temp | 11.7979 | 19.7688 | 15.892 |
| HR | 73.6001 | 38.8815 | 57.9932 |
| BPR | 589.0923 | 588.684 | 588.5526 |
| UVB | 1.8992 | 78.5109 | 2.0361 |
| UVA | 33.9764 | 456.7862 | 28.0838 |
| SO2 | 12.4906 | 6.2314 | 8.9314 |
| Ozone | 13.1753 | 48.4841 | 22.9993 |
| PM10 | 41.4525 | 35.5092 | 45.8033 |

La tabla 5.1 muestra que los niveles de Ozono y temperatura están elevados en las horas 13 y 20.

Tabla5.2.Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 2.

| Hora | 4 | 8 | 16 |
|---|---|---|---|
| WD | 98.4761 | 291.3187 | 155.3659 |
| WS | 2.8427 | 2.9418 | 4.1545 |
| Temp | 13.2001 | 13.0742 | 19.2825 |
| HR | 73.3463 | 72.2136 | 46.6802 |
| BPR | 592.7827 | 593.1286 | 592.1559 |
| UVB | 2.1981 | 4.972 | 42.2521 |
| UVA | 21.6797 | 52.0069 | 223.8936 |
| SO2 | 5.28 | 6.9425 | 3.765 |
| Ozone | 11.9427 | 10.131 | 31.079 |
| PM10 | 30.7329 | 39.2625 | 38.5491 |

En la estación 2 (Tabla 5.2), se presentan los niveles más altos de Ozono y temperatura durante las 16 horas. Los niveles más altos de SO2 y PM10 se muestran en la hora 8.

Tabla 5.3.Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 3.

| Hora | 4 | 13 | 20 |
|---|---|---|---|
| WD | 166.0504 | 178.6099 | 166.7585 |
| WS | 2.2724 | 4.2057 | 3.9289 |
| Temp | 12.0821 | 20.4735 | 16.5192 |
| HR | 74.2233 | 37.1224 | 57.4457 |
| BPR | 591.1377 | 590.4559 | 590.4185 |
| UVB | 3.7514 | 66.6656 | 3.2814 |
| UVA | 33.2962 | 423.4624 | 22.7655 |
| SO2 | 8.2077 | 7.0599 | 7.4112 |
| Ozone | 9.1868 | 34.2708 | 13.357 |
| PM10 | 52.8616 | 44.4201 | 63.4932 |

Los resultados de la estación 3 (Tabla 5.3), se muestran en la tabla 5.3, los niveles altos de Ozono y temperatura están sobre las 13 y 20 horas.

Tabla 5.4. Resultados obtenidos mediante la aplicación de *K-Means* en la estación número 4.

| Hora | 4 | 13 | 20 |
|---|---|---|---|
| WD | 213.8811 | 178.236 | 161.2262 |
| WS | 2.6462 | 4.7285 | 4.0579 |
| Temp | 13.2219 | 21.2551 | 17.3485 |
| HR | 78.6761 | 41.1722 | 61.5275 |
| BPR | 596.7788 | 596.1597 | 596.0998 |
| UVB | 2.696 | 50.8874 | 2.0678 |
| UVA | 30.8936 | 395.6696 | 19.7573 |
| SO2 | 4.4004 | 3.1201 | 2.7674 |
| Ozone | 11.9325 | 44.9311 | 21.5038 |
| PM10 | 53.8374 | 45.2253 | 82.2188 |

En la estación 4 (Tabla 5.4), se aprecia que los niveles más altos de ozono y temperatura se encuentran en las 13 y 20 horas. Los niveles de SO2 se advierten significativos en la mañana y la tarde, así como los niveles de PM10.

Después de explicar los resultados arrojados por la técnica de *K-Means*, se comprueba que los niveles altos de temperatura corresponden con niveles elevados de ozono.

## 5.2 Regresión Lineal

La regresión lineal simple fue usada para construir un modelo del Ozono a partir de los datos meteorológicos, que constituyen sus entradas. Mediante la aplicación de esta técnica, se encontró una relación lineal entre el Ozono y sus atributos (WS, Temperatura, HR, UV-A, UV-B). En la tabla 5.5, se muestran los resultados obtenidos de aplicar la regresión lineal para cada estación de monitoreo y todas las estaciones.

Tabla 5.5. Ecuaciones para predecir los niveles de Ozono de las principales mediciones meteorológicas.

| Número de Ecuación | Estación | Ozono |
|---|---|---|
| 1 | Tecnológico | 0.4569 * WS + 2.1924 * Temp - 0.138 * HR + 0.0112 * UV-A + 0.0847 * UV-B - 5.9446 |
| 2 | Parque de Las Ninfas | 0.1489 * WS + 1.3109 * Temp - 0.1765 * HR + 0.0301 * UV-A + 0.0509 * UV-B + 4.4909 |
| 3 | Serdán | 0.3581 * WS + 0.8397 * Temp - 0.1201 * HR + 0.0204 * UV-A + 0.0698 * UV-B + 5.7526 |
| 4 | Agua Santa | 0.3813 * WS + 1.64* Temp - 0.1871 * HR + 0.0172 * UV-A + 0.0874 * UV-B + 3.0607 |
| 5 | Todas las estaciones | 0.4889 * WS + 1.4666 * Temp - 0.1481 * HR + 0.022 * UV-A + 0.0624 * UV-B + 1.1571 |

De los valores mostrados en la tabla 5.5, el error absoluto promedio fue calculado para todas las funciones. Esto produce un cálculo promedio de 9.405725, lo que significa que los niveles de Ozono tienen un margen de error de +-9.4 ppm.

Mediante la observación de coeficientes de temperatura de las ecuaciones en la tabla 5.5, se confirma que la temperatura contribuye de forma directa a la producción de ozono en el aire. En la ecuación 5, para todas las estaciones de monitoreo, se puede verificar que el

coeficiente de temperatura afecta de manera desproporcionada a la Capa de Ozono, seguido por de la velocidad del viento y la radiación. Sin embargo, HR afecta inversamente a la producción del Ozono.

Tabla 5.6. Comparativo entre los niveles de Ozono reales y las predicciones.

| Estación | Ozono (Regresión) | Ozono (*K-Means*) |
|---|---|---|
| 1 | 48.7352726 | 50.4721 |
| 2 | 42.718172 | 44.5732 |
| 3 | 34.2589031 | 35.3071 |
| 4 | 44.7298418 | 46.1615 |

Se puede observar en la tabla 5.6, que los valores pronosticados de Ozono con una regresión lineal se corresponden adecuadamente, para el promedio de los valores reales de Ozono. Por lo tanto este modelo de regresión puede utilizarse para lanzar alertas de contingencia ambiental.

## 5.3 Árboles de decisión

El primer paso para resolver problemas complejos es descomponerlos en subproblemas más simples. Los árboles de decisión fueron utilizados para proporcionar un modelo base, este modelo tiene como entradas las mediciones de los elementos meteorológicos, la temperatura y la presión entre otros. El resultado arroja un error de +-5ppm con respecto del nivel de Ozono. El modelo resultante es robusto (Figura 5.1), y representa todo un desafío para su interpretación.

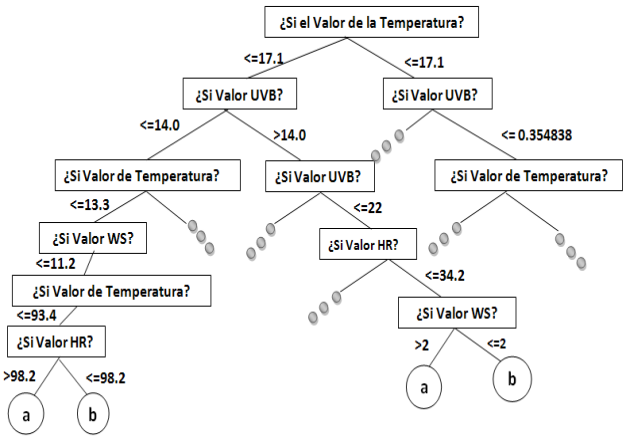



Figura 5.1: Muestra del árbol de decisión obtenido, con respecto de la estación 1.

En el árbol de la figura 5.1, se pueden apreciar una serie de decisiones basadas en las mediciones de temperatura, la radiación UV-B y predicciones de niveles de ozono. La predicción del modelo es de 59,14665% utilizando 10 veces el modo de prueba de validación cruzada. Una vez más la temperatura es considerada como un factor clave meteorológico, estos resultados pueden ser usados para prevenir las contingencias debido a los altos niveles de Ozono en el aire.

## 5.4 Conclusiones

Una vez descritos los resultados obtenidos tras la fase experimental desarrollada en esta Tesis y teniendo en cuenta los objetivos planteados al comienzo de la misma, se pueden enumerar las conclusiones derivadas:

1. Se realizó la búsqueda de cambios en el comportamiento de contaminantes a diferentes horas del día. En conclusión, la técnica de *Clustering*, confirma que los más altos niveles de Ozono son evidentes en horas del medio día (mayor temperatura), mientras que los niveles más altos alcanzados por las partículas de PM10 son relevantes en la noche (20 horas).

2. Se realizó la búsqueda del comportamiento del Ozono con relación a ciertas variables (temperatura). Con la aplicación del algoritmo de K-Medias se encontró que los niveles altos de temperatura se corresponden con niveles altos de ozono.

3. Un modelo lineal fue descubierto en los datos mediante la aplicación de regresión, lo que podría ser utilizado como una herramienta para representar la relación causa-efecto entre el Ozono y las condiciones atmosféricas.

4. Un modelo J48 confirmó que la temperatura es la variable más importante para predecir los niveles de Ozono, seguido de radiación UV-B y el momento en que se obtuvo la medición.

# Capítulo 6

# Conclusiones y Trabajo a Futuro

En este capítulo se presentan las conclusiones de la investigación a través de tres secciones dedicadas a: exponer una discusión sobre las lecciones aprendidas durante el desarrollo de la tesis; precisar las aportaciones alcanzadas; y presentar las líneas de investigación futuras que se derivan de los resultados arrojados por la Tesis.

## 6.1 Aportaciones

Como aportaciones de esta Tesis, se podrían citar las siguientes:

- Se generaron modelos que establecen que el nivel de Ozono tiene una dependencia directa con la temperatura (ésta en mayor proporción), producto de la aplicación de Técnicas de Minería de Datos: *K-Means*, Regresión Lineal y J48.
- Se confirmó la confiabilidad de la información generada por la SEMARNAT.
- Los resultados que brinda la Minería de Datos en conjunto con expertos en el tema, permiten construir como en este caso, modelos predictivos que son necesarios para la creación de planes estratégicos para contingencias relacionadas con el incremento de los contaminantes en el aire.

## 6.2 Líneas de Investigación Futuras

Las nuevas líneas de investigación que tras el desarrollo de esta Tesis quedan aún abiertas, en el campo del “Cambio Climático y la toma de decisiones” son los siguientes:

- Se está trabajando en el desarrollo de técnicas para predecir los niveles de otros contaminantes como las partículas PM10 y SO2 con modelos no lineales, ya que preliminarmente no se observa una relación lineal entre estos y las condiciones climáticas.

- Del punto anterior, se plantea identificar las patologías para el diagnóstico de ciertas enfermedades relacionadas con las partículas $PM_{10}$ . Así como también permitir la detección de grupos de población con riesgo de sufrir una patología concreta.

- Con los resultados obtenidos de este trabajo de Tesis, se plantea la construcción de un Manual de Contaminantes Ambientales, para emitir alarmas a la población y así tomar las medidas necesarias en ciertas épocas del año.

- Se plantea lanzar campañas organizacionales más adecuadas para atenuar el Cambio Climático.

## 6.3 Conclusiones Finales

En esencia, la Tesis se encaminó a proponer el análisis, diseño e implementación de un sistema para la toma de decisiones con respecto del Cambio Climático.

La experiencia obtenida en conjunto con la Secretaria del Medio Ambiente y Recursos Naturales del Estado de Puebla, en la realización de este trabajo, permite concluir que es factible la construcción de sistemas para la toma de decisiones basados en la perspectiva de los siguientes puntos:

a) Existe una gran cantidad de información que actualmente no está siendo aprovechada en toda su dimensión.

b) Existe un software de Minería de Datos de distribución libre y gratuita, fácil de usar y que contiene herramientas necesarias para el análisis.
c) Este software de Minería de Datos puede ser utilizado por una persona ajena al ámbito informático con una capacitación básica.
d) La unión de expertos en materia de Medio Ambiente con especialistas en el área de computación ha sido de gran relevancia en cuanto a los resultados obtenidos de esta Tesis.

# Referencias


1. José C. Riquelme, Roberto Ruiz y Karina Gilbert. Minería de Datos: Conceptos y Tendencias. Revista Iberoamericana de Inteligencia Artificial, 10 (029):11-18, Diciembre 2006.

2. José Hernández Orallo, M. José. Ramírez Quintana y César Ferri Ramírez. Introducción a la Minería de Datos. Reimp. Madrid España. Pearson. 2004. 656p.

3. Instituto Nacional de Ecología [en línea]: Naucalpan, Edo. México. Dirección de Investigación en Monitoreo Atmosférico y Caracterización Analítica de Contaminantes. Marzo 2005. [fecha de consulta: Marzo 2010]. Disponible en: <http://sinaica.ine.gob.mx/>.

4. Julia Martínez, Adrián Fernández Bremauntz y Patricia Osnaya. Cambio Climático: una visión desde México. Noviembre de 2004. México D.F.

5. DataPrix, Introducción al Manual de DataWarehouse [en línea]: Edo. México. DataPrix [fecha de consulta: Marzo 2010]. Disponible en: <http://www.dataprix.com/introduccion-manual-dwh>

6. Angoitia Espinoza, Itziar. DataWarehouse para la gestión de lista de espera sanitaria. Tesis (Licenciatura en Informática). Madrid España, Facultad de Informática Universidad Politécnica de Madrid. 2008. 135p.

7. Marysol Tamayo y Francisco Javier Moreno. Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP: Comparing the MOLAP the ROLAP storage models. Revista Ingeniería e investigación. 26(003): 135-142, 2006.

8. Datawarehousing: Teoría.Introducción al Concepto DataWarehousing [en línea]: España. [fecha de consulta: Marzo 2010]. Disponible en : <http://personal.lobocom.es/claudio/gen006.htm>

9. Chaudhuri, 1997; Kimball, 2002.

10. Yan Zhu, Christof Bornhövd, Doris Sautner, and Alejando P. Buchmann. Materializing Web Data for OLAP and DSS. Department of Computer Science, Darmstadt University of Technlology. 1846/2000 :( 201-215), 2000.

11. Rosette Uzcanga, José Miguel Dominique. Sistema de apoyo a la toma de decisiones a partir de documentos distribuidos en el Web: aplicación a la prensa electrónica.Tesis

(Licenciatura Ingeniería en Sistemas Computacionales). Puebla, Puebla. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas Puebla. 2003.

12. ETL-Tools.Info. Esquema de constelación de hechos (fact constellation shema) [en línea]: ETL-Tools 2006-2010 [fecha de consulta: Marzo 2010]. Disponible en: < http://etl-tools.info/es/bi/almacendedatos_esquema-constelacion.htm>

13. Sinnexus Business Intelligente + Informática estratégica [en línea]: Sinnexus-Ronda de Outeiro No 116-15008, 2007 [fecha de consulta: Marzo 2010]. Disponible en : <http://www.sinnexus.com/business_intelligence/datamart.aspx>

14. José Hernández Orallo. Parte II: Almacenes de Datos, transparencias basadas parcialmente en el "tutorial DW" de Matilde Celma [diapositiva]. Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. 122 diapositivas, col.

15. Isabel Criado Gómez y Paloma Sánchez López. OLAP, ROLAP, MOLAP [en línea]: Ministerio de Trabajo y Asuntos Sociales, 2006 [fecha de consulta: Marzo 2010]. Disponible en: <http://www.csi.map.es/csi/silice/DW2251.html#ROLAP>

16. Creative Commons Reconocimiento Compartir Igual 3.0. MOLAP [en línea]: Julio 2010 [fecha de Consulta: Marzo 2010]. Disponible en: <http://es.wikipedia.org/wiki/MOLAP>

17. Kimball 1996

18. NewTec Ediciones, 2002.

19. Ibarra, 2005.

20. Nader, 2003.

21. Inmon 2002.

22. Creative Commons Reconocimiento Compartir Igual 3.0. Extract, transforma and load [en línea]: Agosto de 2010. [fecha de consulta: Abril 2010].Disponible en <http://es.wikipedia.org/wiki/Extract,_transform_and_load>

23. José Hernández Orallo, M. Carmen Juan Lizandra, Neus Minaya Collado y Carlos Monserrrat Aranda. Extracción y Visualización de Conocimiento de Bases de Datos Médicas. ACTA, Vol. 18, 49-58, 2000.

24. Apolinar Velarde Martínez. Sistema de visión artificial para la verificación del llenado de recipientes no opacos utilizando redes neuronales artificiales. Tesis (Maestría en Desarrollo). México, Centro Nacional de Investigación y Desarrollo Tecnológico. Julio de 1998.

25. Thornton 2000.

26. Fayyad, Piatetsky-Shapiro y Smyth 1996.

27. Moody & Darken 1989, MacQueen 1967.

28. José C. Riquelme, Roberto Ruiz y Karina Gilbert. Minerìa de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, primavera. 10(029): 11-18, 2006.

29. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy. Advances in Knowledge Discovery and Data Mining. IEEE Expert.11(14):20-25, Octubre 1996.

30. Fayyad et al. 1996a.

31. Moody & Darken 1989, MacQueen 1967.

32. Nikhil R. Pal and Lakhmi Jain (Eds). Advanced Techniques in Knowledge Discovery and Data Mining. London. Springer-Verlarg.2005. 234 p.

33. Monitoring Atmospheric Composition & Climate. European Air Quality [en línea]: 2009-2011. [fecha de consulta: Abril 2010]. Disponible en : < http://www.gmes-atmosphere.eu/services/raq/raq_nrt/>

34. Cebtri de Previsai de Tempo e Estudos Climáticos. CPTEC [en línea]: 1995-2010. [fecha de consulta: Abril 2010]. Disponible en: < http://www.cptec.inpe.br/> .

35. Centro del Agua del Trópico Húmedo para América Latina y El Caribe-CATHALAC. Fomento de las Capacidades para la Etapa II de Adaptación al Cambio Climático en Centroamérica, México y Cuba [en línea]: 2008. [fecha de consulta: Mayo 2010]. Diponible en: < http://www.cathalac.org/Publicaciones/Cambio-Climatico/Fomento-de-las-Capacidades-para-la-Etapa-II-de-Adaptacion-al-Cambio-Climatico>

36. Dr. Víctor O. Magaña y Dr. Carlos Gay García. Vulnerabilidad y Adaptación Regional ante el Cambio Climático y sus Impactos Ambiental, Social y Económicos. Gaceta Ecológica. (065): 7-13, Octubre-Diciembre 2002.