UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

LA MINERÍA DE DATOS COMO HERRAMIENTA PARA LA TOMA DE DECISIONES EN EL PROCESO DE CALENDARIZACIÓN DE CURSOS DE CÓMPUTO

TESIS PROFESIONAL
QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN INFORMÁTICA

PRESENTA:

CARLOS TOMÁS REYES GARCÍA

ASESOR:

L.I. CARLOS FRANCISCO MÉNDEZ CRUZ

MÉXICO, D.F. **2007**

Para mis padres: Tomás y Soledad.

Gracias por darme una gran educación y guiarme por la vida en los momentos difíciles. Sin ustedes nada habría sido igual. Los amo.

A mi hermano Jorge.

Gracias por el apoyo incondicional, por ser tú mismo.

A mis demás hermanos: Arabia, Carlos, Darío, Fernando, Rocío y Teresita.

Con ustedes he aprendido que no sé necesita llevar la misma sangre para poder decir te quiero, carnal. Gracias por todo el apoyo, por las risas y abrazos.

A Elvira Salgado.

Gracias por existir, por enseñarme que la vida esta llena de color.
Te amo.

A Carlos Méndez.

Gracias por los consejos, por darme la oportunidad de conocer un poco más de ti.
“Las palabras precisas de un maestro siempre van a determinar el camino de un alumno.”

Contenido

Contenido I

Introducción 3

Objetivos 5

Capítulo 1

Evolución de las Bases de Datos 7

1.1. La Necesidad Histórica de Almacenar Información 7

1.2. Bases de Datos: Definición y Conceptos 10

1.2.1. Dato 11

1.2.2. Banco de Datos 11

1.2.3. Campo 11

1.2.4. Registro 11

1.2.5. Archivo 11

1.2.6. Información 12

1.2.7. Sistema de Información 12

1.2.8. Sistema de Bases de Datos 12

1.2.8.1. Objetivos de un Sistema de Bases de Datos 12

1.2.9. Sistema Manejador de Bases de Datos 13

1.3. Modelos de Bases de Datos 13

1.3.1. Organización Física de los Datos 13

1.3.1.1. Medios de Almacenamiento 13

1.3.2. Modelo de Archivos 14

1.3.2.1. Modelo Jerárquico 16

1.3.2.2. Modelo de Red 17

1.3.3. Modelo Relacional 18

1.3.4. Modelo Orientado a Objetos 21

1.4. Necesidades actuales 22

1.5. ¿Qué es la Minería de Datos? 23

Capítulo 2

Minería de Datos 25

2.1. Definición 25

2.2. Modelos de Minería de Datos 26

2.2.1. Modelo Deductivo 26

2.2.2. Modelo Predictivo 27

2.3. Funcionalidades de la Minería de Datos 27
2.3.1. Descripción de Clases/Conceptos: Caracterización y Discriminación 28
2.3.1.1. Caracterización 28
2.3.1.2. Discriminación 28
2.3.2. Estimación 29
2.3.3. Clasificación 30
2.3.4. Agrupamiento (Clustering) 30
2.3.5. Predicción 31
2.3.6. Asociación 32

Capítulo 3

Metodología para Minería de Datos: CRISP – DM 1.0 33
3.1. CRoss – Industry Standard Process of Data Mining 33
3.2. Comprensión del negocio 36
3.2.1. Determinación de los objetivos del negocio 36
3.2.2. Evaluación de la situación 37
3.2.3. Determinación de objetivos en términos de Minería de Datos 38
3.2.4. Planeación del proyecto 38
3.3. Comprensión de los datos 39
3.3.1. Compilación inicial de los datos 39
3.3.2. Descripción de los datos 39
3.3.3. Exploración de los datos 40
3.3.4. Verificación de los datos 40
3.4. Preparación de los datos 41
3.4.1 Selección de datos 41
3.4.2. Limpieza de datos 42
3.4.3. Construcción de datos 42
3.4.4. Integración de datos 43
3.4.5. Formateo de datos 43
3.5. Modelado 44
3.5.1. Elección del modelo 44
3.5.2. Diseño de pruebas 44
3.5.3. Construcción del modelo 45
3.5.4. Evaluación del modelo 45
3.6. Evaluación 46
3.6.1. Evaluación de resultados 46
3.6.2. Reseña del proceso 47
3.6.3. Asignación de nuevas tareas 47
3.7. Despliegue 48
3.7.1. Plan de despliegue 48

3.7.2. Plan de monitoreo y mantenimiento 48
3.7.3. Elaboración del reporte final 49
3.7.4. Reseña del proyecto 49

Capítulo 4

Enfoque práctico de la Minería de Datos 51
4.1. La Minería de Datos en los negocios 51
4.2. El rol de los sistemas de información operacionales 52
4.3. El rol del Data Warehouse 52

Capítulo 5

Minería de Datos: Aplicación en un caso práctico 55
5.1. Comprensión del negocio 55
5.1.1. Determinación de los objetivos del negocio 56
5.1.1.1. Objetivos organizacionales 56
5.1.1.2. Perspectiva organizacional sobre los resultados esperados 56
5.1.2. Evaluación de la situación 57
5.1.2.1. Inventario de recursos 57
5.1.2.2. Requerimientos, suposiciones y restricciones 57
5.1.2.3. Convenio de confidencialidad de la información a utilizar 58
5.1.2.4. Aprobación sobre la publicación del proyecto 60
5.1.3. Determinación de objetivos en términos de Minería de Datos 61
5.1.3.1. Objetivo en términos de Minería de Datos 61
5.1.3.2. Criterio de éxito 61
5.1.4. Planeación del proyecto 61
5.1.4.1. Alcance 61
5.1.4.2. Objetivo del proyecto 61
5.1.4.3. Actividades a realizar 62
5.1.4.4. Herramientas a utilizar 63
5.2. Comprensión de los datos 63
5.2.1. Compilación inicial de los datos 63
5.2.1.1. Reporte inicial sobre la adquisición de los datos 63
5.2.2. Descripción de los datos 64
5.2.2.1. Reporte sobre la descripción de los datos 64
5.2.3. Exploración de los datos 65
5.2.3.1. Reporte de la exploración de los datos 65
5.2.4. Verificación de los datos 65
5.2.4.1. Reporte de la calidad de los datos 65
5.3. Preparación de los datos 66
5.3.1. Selección de datos 66
5.3.2. Limpieza de datos 67
5.3.2.1. Reporte sobre la limpieza de datos 67

5.3.3. Construcción de datos 67
5.3.3.1. Atributos derivados 67
5.3.4. Integración de datos 68
5.3.4.1. Reporte sobre la integración de datos 68
5.3.5. Formateo de datos 69
5.3.5.1. Conjunto de datos con formato 69
3.3.6. Resultados de la fase 70
5.4. Modelado 71
5.4.1. Elección del modelo 71
5.4.1.1. Definición del modelo a utilizar en el proyecto 71
5.4.2. Análisis de Componentes Principales 72
5.4.2.1. La nube de puntos fila 72
5.4.2.2. La nube de puntos columna 73
5.4.2.3. Resultados del ACP 74
5.4.2.4. Criterios para seleccionar el número de componentes 75
5.4.3. Diseño de pruebas 76
5.4.3.1. Documentación de pruebas 76
5.4.4. Construcción del modelo 77
5.4.4.1. Descripción del modelo 77
5.4.5. Evaluación del modelo 79
5.4.5.1. Reporte sobre la evaluación del modelo 79
5.4.5.2. Revisión de los parámetros 79
5.5. Evaluación 79
5.5.1. Evaluación de resultados por los especialistas en Minería de Datos 79
5.5.2. Evaluación de resultados por los dueños del negocio 90
5.5.3. Reseña del proceso 91
5.6. Despliegue 92
5.6.1. Plan de despliegue 92
5.6.2. Plan de monitoreo y mantenimiento 93
5.6.3. Elaboración del reporte final 94

Capítulo 6

Conclusiones 97

Bibliografía 99

Introducción

Las organizaciones de hoy día, no importa si son pequeñas, medianas o de gran envergadura, se mantienen en constante evolución. Así, la información se ha convertido en el activo más valioso para una organización. Vivir en un mundo globalizado obliga a las organizaciones a establecer estrategias enfocadas a obtener una mayor ventaja competitiva dentro su campo de aplicación; en este sentido las tecnologías de la información han venido siendo, desde la última década, la herramienta base para el soporte operacional de una organización.

Los sistemas de información son pieza fundamental para las organizaciones, ya que, ellos proporcionan la base de conocimiento más importante que una organización pueda tener, así mismo, los sistemas de información se estructuran de forma tal que se genere información estratégica que dé pauta a realizar una adecuada toma de decisiones.

Las tecnologías de la información constituyen por tanto parte fundamental de una organización, ya que mediante el uso de estas tecnologías se realiza la administración del activo más valioso.

Así pues, luego de la puesta en marcha de una organización, no pasará mucho tiempo para que ésta tenga almacenes de datos con un gran volumen. La visión que empieza a tener una organización sobre grandes volúmenes de datos comienza a ser más abstracta, dejando de lado el detalle y los beneficios que ello conlleva.

El correcto uso de la información por parte de una organización, es considerado un factor crítico de éxito, ya que dicha información representa la historia de una organización, historia de la cual se aprende para generar el conocimiento que permita su evolución.

Como se planteó anteriormente, cuando las organizaciones cuentan con grandes volúmenes de información, la visión que de ella hacen comienza a ser cada vez más abstracta, así pues, en este documento de tesis se plantean los beneficios de explotar grandes volúmenes de datos a través de técnicas estadísticas, matemáticas y computacionales, las cuales permitan la visualización de patrones ocultos en los datos, patrones que son traducidos en conocimiento nuevo, útil y entendible para las organizaciones.

Hallar conocimiento nuevo dentro de grandes volúmenes de datos permitirá a las organizaciones obtener una ventaja competitiva adicional dentro de su campo de aplicación.

Este documento de tesis nace por la necesidad de dar explicación al fenómeno de impartición de cursos de cómputo, en donde a partir de un gran volumen de datos se hagan visibles patrones ocultos que enriquezcan la visión del modelo de negocio de una institución académica como lo es la Dirección General de Servicios de Cómputo Académico de la UNAM.

El objetivo principal de este documento de tesis es responder a la pregunta de si ¿la Minería de Datos puede ser utilizada como herramienta para la toma de decisiones en el proceso de calendarización de cursos de cómputo?

Bajo el marco de referencia anterior es como se plantea a la Minería de Datos como objeto de estudio de este documento de tesis.

A lo largo de los siguientes capítulos, se encontrarán los fundamentos de Minería de Datos, los cuales proponen a través de diversas técnicas la generación de conocimiento proveniente de grandes volúmenes de datos.

También, se ha hecho énfasis sobre el uso de una metodología para el seguimiento de un proceso de Minería de Datos, el capítulo tercero de este documento de tesis documenta el proceso que sigue un proyecto de Minería de Datos desde el enfoque de la metodología CRSIP-DM, se evalúa cada una de las fases consistentes y un subconjunto de tareas específicas propuestas.

En capítulos posteriores se da el enfoque de la Minería Datos desde la visión de los negocios, culminando con un caso de aplicación práctico el cual refleja la puesta en marcha del aprendizaje obtenido luego de evaluar a la Minería de Datos desde un marco teórico.

El caso de aplicación práctico se enfoca en la evaluación de la hipótesis inicial y por ende sirve para ilustrar cuestiones técnicas que en capítulos previos se plantean, así mismo refleja el uso de la metodología de Minería de Datos antes propuesta.

Sin más preámbulos sírvase a comenzar con la lectura de este material, que ha sido desarrollado para llevar al lector de la mano teniendo como propósito principal sentar las bases de la Minería de Datos con una prospección hacia su profundización en documentos técnicos más específicos.

Objetivos

A lo largo del presente documento de tesis se realizará: una investigación sobre los diversos elementos que conforman a la Minería de Datos y la incursión de un caso práctico, mediante los cuales sea posible alcanzar los siguientes objetivos:

- Principal.

  Determinar si la Minería de Datos puede ser utilizada como una herramienta para la toma de decisiones en el proceso de calendarización de cursos de cómputo.

- Secundarios.

  Proveer de un sistema semiautomático de Minería de Datos a la Dirección de Cómputo para la Docencia de la DGSCA UNAM, el cual le permita obtener una categorización de cursos de cómputo basada en la efectividad de los mismos.

  Sentar las bases para la realización de análisis ulteriores de Minería de Datos dentro del modelo de negocio que sigue la Dirección de Cómputo para la Docencia de la DGSCA, UNAM.

- Académicamente el trabajo presentado tiene como objetivo principal que el lector sea capaz de:
  - Adquirir una visión global e integrada del los aspectos relativos a la Minería de Datos (Antecedentes, Modelos, Técnicas, Herramientas, Metodología, etc.)
  - Comprender la importancia del uso de una metodología para la generación de conocimiento dentro de un proceso de Minería de Datos.

Capítulo 1

Evolución de las Bases de Datos

Las Bases de Datos representan una herramienta necesaria en la vida de las personas, comprender su evolución y funcionamiento será fundamental para obtener un panorama más amplio de la situación actual y tomar decisiones que encaminen hacia un futuro promisorio. Así pues, este capítulo tiene como objetivo presentar al lector los componentes que históricamente han integrado a las Bases de Datos, su funcionamiento y nuevos retos ante el creciente desarrollo tecnológico y necesidades de almacenar información.

1.1. La Necesidad Histórica de Almacenar Información

El hombre desde que comienza a llevar una vida sedentaria descubre la gran necesidad de formar grupos con el objeto de facilitar las tareas que conlleva una vida en sociedad; las tareas y trabajos que realizaban los primeros grupos eran básicamente los necesarios para sobrevivir, requirió de mucho tiempo especializarse en ciertas actividades y, conocer el funcionamiento de las cosas fue en principio difícil, sin embargo, para que el grupo perdurará debían hallar la forma de transmitir toda la experiencia y conocimiento adquirido.

En principio, el conocimiento era transmitido generación tras generación a través de la enseñanza directa, los padres enseñaban a sus hijos las técnicas aprendidas durante su vida con el objeto de preservar el grupo y asegurar la supervivencia de los suyos. Más tarde comenzaron a aparecer diversas expresiones artísticas como lo son las pinturas rupestres cuyo objetivo era plasmar parte de la vida del hombre, en estas pinturas se relataban historias, acontecimientos que se deseaba perduraran para ser apreciados por generaciones futuras; así pues, nos encontrábamos ante el primer banco de datos de la historia del hombre, las pinturas rupestres, en dónde los primeros hombres plasmaban acontecimientos de su vida con el objeto de transmitir la experiencia adquirida a generaciones futuras.

Conforme los grupos fueron creciendo y expandiéndose a lo largo y ancho del planeta, fueron constituyéndose las primeras civilizaciones del mundo y con ello se constituyeron también los primeros sistemas de comunicación especializados, en dónde no sólo existían expresiones orales o artísticas, sino que, expresiones como la escritura a través de tablillas de barro ó códices ya formaban parte de los sistemas de comunicación de éstas primeras civilizaciones.

En la antigua Mesopotamia las primeras expresiones de lenguaje escrito eran plasmadas en tablillas de barro que eran esculpidas con el objeto de preservar la historia y conocimiento.

Grandes civilizaciones de Mesoamérica como los mayas, aztecas, mixtecos, zapotecas, otomíes y purépechas, entre otros, registraron sus conocimientos en los códices (*lat.* codex, "libro manuscrito")

desde épocas muy remotas; la información que éstos proporcionan permite apreciar los diversos aspectos culturales, sociales, económicos y científicos desarrollados por los pueblos antiguos, como sus creencias religiosas, ritos, ceremonias, nociones geográficas, historia, genealogías y alianzas entre los señoríos, sistema económico y cronología.

La evolución del hombre, en gran medida debido a la transmisión de conocimiento permitió que surgiesen nuevas técnicas para almacenar la información generada, los egipcios escribían sobre papiro el cual se obtenía a partir de una caña abundante en el Río Nilo. (Web. 01)

Hasta este punto de la historia del hombre podemos decir con certeza que habían surgido las primeras bases de datos de la historia, claro que con algunas restricciones, pero con la característica primordial que es la de almacenar información y recuperarla en un futuro con el objeto de transmitir conocimiento que permitiese la evolución de la especie.

En Europa, durante la Edad Media, se utilizó el pergamino que consistía en pieles de cabras y carneros preparadas para recibir la tinta, por desgracia eran bastante caros, lo que ocasionó que a partir del siglo VIII se popularizara la infausta costumbre de borrar los textos de los pergaminos para escribir sobre ellos otros distintos, perdiéndose de esta manera una cantidad inestimable de obras. Sin embargo, los chinos ya fabricaban papel a partir de los desechos de la seda y el cáñamo e incluso del algodón, invento que se atribuye a Cai Lun, debido a la relación comercial trasmitieron este conocimiento a los árabes, quienes a su vez lo llevaron a lo que hoy son España y Sicilia desde el siglo X. La elaboración de papel se extendió a Francia que lo producía utilizando lino desde el siglo XII. (Web. 01)

Los avances tecnológicos de la civilización china permiten la creación del papel, éste surge debido a la necesidad de almacenar volúmenes más grandes de información, ya que la utilización de pergaminos resultaba costosa además de estar limitada. El papel por lo tanto se convierte en la principal fuente de almacenaje de información, y de igual forma que en épocas más antiguas en éste se plasmaba el conocimiento, la historia, las costumbres y creencias.

El papel era tan sólo el reflejo de la evolución tecnológica que la humanidad había alcanzado, porque no sólo era el papel sino que ya existían lenguajes idiomáticos con mayor formalidad, se habían constituido pueblos, imperios y posteriormente naciones; existían diferentes ámbitos socio-culturales que necesitaban del papel y la tinta para plasmar ideas en lenguajes diversos. Transmitir el conocimiento empezaba a dificultarse, el hombre había creado barreras idiomáticas, además de las barreras geográficas entendibles, y si a esto sumamos que la reproducción de obras debía hacerse de forma manual, obtenemos como resultado una desaceleración en el proceso de evolución; el conocimiento podía duplicarse o no tomarse como base para generar nuevas ideas, esto debido a que la difusión de obras no podía presentarse en el momento justo.

Para la generación de libros en un principio se debía diseñar manualmente página a página, lo que hacia de la reproducción de obras un proceso lento ya que de existir algún error debía comenzarse de nuevo; y no fue sino hacia el año de 1450 cuando el alemán Johannes Gutenberg inventa la imprenta, un conjunto de mecanismos que permitían la impresión en papel y que en esencia representaba un conjunto de juegos tipográficos fabricados en un principio en madera y más tarde en plomo. Con la invención de la imprenta se satisfacía la necesidad de difusión de obras literarias, científicas e informativas como lo es el periódico.

La imprenta fue evolucionando a pasos agigantados, la impresión de libros fue haciéndose constante a cada momento, se comenzó con la construcción de grandes bancos de datos como lo son las bibliotecas, en dónde se almacenaban copias de las obras impresas y que podían ser consultadas en el instante preciso, Asimismo con la evolución de las comunicaciones y transportes, el tiempo que tardaba una obra en distribuirse en el mundo comenzaba a ser menor aunque aún muy tardado.

Por un lado la impresión de libros permitía una mayor difusión de contenidos, Asimismo estos podían consultarse en bibliotecas, por otra parte aún se contaba con una gran desventaja ya que las actualizaciones o modificaciones hacia los contenidos era un proceso lento debido a que debía reimprimirse una edición y esto significaba además, un elevado costo, ya que no era palpable la modificación sino hasta la redistribución de los ejemplares, cabe destacar también que las cantidades de papel acumulado en las bibliotecas aumentaba de forma vertiginosa. En la era moderna, ya situados en una época más actual, los volúmenes de papel eran enormes y realizar una búsqueda sobre algún tema específico tomaba grandes cantidades de tiempo, dinero y esfuerzo.

A través de este capítulo hemos podido darnos cuenta de la gran necesidad del hombre por conservar su historia, ya que, ésta le enseña los errores que ha cometido, y el conocimiento adquirido, que son principalmente la base de la evolución. Conservar la historia significa que el hombre se ha preocupado por almacenarla y resguardarla de tal forma que esa enseñanza alojada pueda ser transmitida hacia generaciones futuras con el objeto de redimensionarla para dar un salto en la evolución.

Es bueno señalar que se ha dejado de lado gran parte de la historia de la humanidad ya que no es el objetivo principal del tema, sino que, se han realizado saltos en la historia del hombre con el objeto de señalar de forma precisa aquellos acontecimientos que han marcado los antecedentes de las Bases de Datos: las pinturas rupestres, las tablillas de barro, los códices e inscripciones jeroglíficas en piedra, los pergaminos, los papiros y el papel; todos estos medios para almacenar información a excepción de las pinturas rupestres quizá, estaban organizados bajo un modelo, es decir, la forma de almacenar la información debía hacerse de cierto modo, por ejemplo para las tablillas de barro, éstas debían ser talladas o inscritas con herramientas especificas y una vez terminadas debían ser presentadas y preservadas con técnicas especiales; para el papel debía utilizarse tinta especial y la forma de almacenar la información era en bibliotecas, dónde ésta, agrupa los libros bajo normas especiales.

El entorno bajo el cuál se desenvuelve el hombre delimita su forma de actuar, es una regla natural, a lo largo de la historia el entorno bajo el cuál se ha desenvuelto el hombre a sufrido grandes modificaciones, la tecnología creada por el hombre es la responsable ya que siempre se ha buscado la forma de realizar las tareas con menor esfuerzo. Los grandes avances tecnológicos de la edad media, la edad moderna y gran parte de la edad contemporánea han sido precedidos por revoluciones industriales, que han marcado la transición entre una y otra edad, pues en estas revoluciones han surgido los elementos necesarios para dar un gran salto en la evolución y transformación del entorno del hombre.

A mediados del siglo XX y sin reconocerse de manera formal es cuando aparece la más reciente revolución industrial de la historia, el nacimiento de los circuitos integrados da pié a este gran acontecimiento, desde ese momento el hombre ha experimentado un acelerado proceso de evolución tecnológica ya que se sienta el precedente de las telecomunicaciones y las computadoras, elementos que sin duda hoy día intervienen en la vida diaria de las personas. Hasta antes de este punto el hombre tenía como medio para almacenar la información los libros, contenidos en las grandes bibliotecas del mundo, sin embargo, la computadora sienta un gran precedente en la historia de la humanidad ya que cambiará el paradigma del libro.

Las computadoras son introducidas en la historia de la humanidad a finales de la década de 1940 y desde ese entonces han sufrido una gran y rápida evolución, motivada principalmente por fines bélicos en donde se desea procesar y transmitir información de forma eficiente y rápida, así las computadoras pasan de ser un gran conjunto de mecanismos que abarcan hasta una sala entera a ser componentes de uso personal.

Las primeras computadoras únicamente eran capaces de realizar cálculos matemáticos a grandes velocidades, sin embargo poco a poco se fueron allegando de características que les permitían

almacenar pequeñas porciones de información, esto es permitido por medios de almacenamiento que a su vez fueron evolucionando también.

La evolución conjunta, tanto de las computadoras como de los medios de almacenamiento, permitieron generar equipos de cómputo con capacidades mayores, menor tamaño, mayor velocidad y mayor capacidad de almacenamiento.

Los primeros medios de almacenamiento utilizados por una computadora, son las memorias, que permiten almacenar información temporalmente hasta que la computadora es apagada o reiniciada, más a delante en la historia de las computadoras, aparecen los medios magnéticos para almacenar información, dichos medios magnéticos tienen la característica de almacenar información y poderla recuperar en un futuro por medio de mecanismos lectores.

Asimismo, dentro de la evolución tecnológica de las computadoras aparecen las redes de computadoras, por sí sola una computadora puede almacenar un buen volumen de información, pero como se ha visto a lo largo de la historia del hombre, éste requiere transmitir la información a los demás para evolucionar, entonces una computadora por sí sola no tenía esa característica y lo que hace el hombre para satisfacer esa necesidad es generar un conjunto de mecanismos que permiten conectar a un grupo de computadoras en red y esto desencadena en que la información puede estar siendo compartida, más allá de las barreras físicas y geográficas entendibles.

El hombre de hoy no sólo desea almacenar el conocimiento generado individualmente, también ha creado una infinidad de organizaciones que persiguen fines particulares, estas organizaciones también cuentan con una historia y al hombre le interesa almacenar la información respecto de ella, con el objeto de procesarla y utilizarla para satisfacer los fines que persigue, no sólo los medios de almacenamiento se vuelven importantes en este punto de la historia, sino que, también requiere de mecanismos que le faciliten la administración de la información. Para realizar esta administración de la información el hombre utiliza como herramienta la computadora ya que le va a permitir simplificar tareas que manualmente tomarían mayor cantidad de tiempo, utilizar la computadora para este fin involucra hacer uso de sistemas de bases de datos los cales están pensados para el manejo de grandes volúmenes de información.

Bajo este contexto hemos explicado las diversas razones por la cuales el hombre decide almacenar la información que genera, principalmente porque le interesa conservar su historia para comprender de dónde viene y hacia dónde va, en este sentido la evolución del hombre en gran medida se debe a que ha sabido conservar y transmitir el conocimiento que genera, hasta este punto es como tenemos completo el marco de referencia bajo el cuál comprendemos “La Necesidad Histórica de Almacenar Información”, así como los distintos medios bajo los cuáles se ha satisfecho esta necesidad.

En temas siguientes exploraremos los sistemas de bases de datos, que en la actualidad son los mecanismos que se utilizan para administrar la información que el hombre genera, así como los medios de los que se vale para el almacenamiento de la misma.

1.2. Bases de Datos: Definición y Conceptos

Ahora que se tiene presente cuál es el marco de referencia bajo el cuál se han desenvuelto las bases de datos en la historia, es pertinente definir con claridad los elementos que intervienen en un sistema de bases de datos y demás conceptos necesarios para adentrarnos en temas posteriores.

1.2.1. Dato

Dato es un conjunto de caracteres con algún significado, pueden ser numéricos, alfabéticos o alfanuméricos. Un dato por sí solo no refleja información útil, puesto que si no se encuentra bajo un contexto es difícil realizar una interpretación de él.

Ejemplos de dato:{1500, H, cancelado, REGC850121}.

1.2.2. Banco de Datos

Sin menor duda es la definición que más concierne a este capítulo, un banco o base de datos puede definirse en principio como *"Una colección de datos almacenados y organizados de alguna forma y con un fin determinado"*.

Sobre esta definición que resulta muy escueta en principio vale la pena hacer algunas observaciones con el objeto de madurar la definición inicial.

Cuando afirmamos que es una colección de datos almacenados necesariamente debemos referirnos a un concepto fundamental en las bases de datos: *Persistencia*, datos que pueden ser recuperados en el futuro independientemente del programa que los manipula.

A su vez, en la definición inicial mencionamos que los datos deben estar organizados de alguna forma, esto debido a que existen diversos *Modelos de Bases de Datos*, y es necesario que nos ajustemos a alguno de ellos, en los temas siguientes revisaremos con más detenimiento dichos modelos.

La tercera parte de la definición inicial menciona que se debe perseguir un fin determinado, éste varía de acuerdo al modelo de negocio de la organización que implementa el sistema de base de datos, sin embargo, dicho fin determinado puede en algún momento ser: *Tomar decisiones, Realizar consultas generales, Evaluar estados financieros, Analizar ventas estadísticas etc.*

Así pues, podemos redefinir el concepto inicial para concluir que Base de Datos es un conjunto de datos persistentes, organizados bajo un modelo y que son utilizados por un sistema de información para la toma de decisiones.

1.2.3. Campo

Espacio mínimo de almacenamiento en un archivo.

1.2.4. Registro

Conjunto de campos que guardan relación entre ellos

1.2.5. Archivo

Unidad de almacenamiento en medios magnéticos. Espacio de almacenamiento en memoria. Algunos conceptos relacionados son:

- BOF (Begin Of File), EOF (End Of File).
- Métodos de acceso.
  - Secuencial e indexado.
- Tipos.

- Texto-planos-ASCII.
- Binarios.

Estos conceptos se revisarán a fondo en el tema 1.3.1 Modelo de Archivos.

1.2.6. Información

Información es un conjunto de datos contextualizados que proporcionan un significado útil; también es vista como el activo más valioso de una organización ya que le va a permitir tomar decisiones que la encaminen hacia los objetivos planteados en la misión organizacional.

1.2.7. Sistema de Información

Un Sistema de Información es un conjunto de elementos de software y hardware que realizan funciones para el procesamiento, distribución y presentación de la información con el objetivo primordial que es la toma de decisiones.

1.2.8. Sistema de Bases de Datos

Los sistemas de base de datos se diseñan para manejar grandes volúmenes de información, la manipulación de los datos involucra tanto la definición de estructuras para el almacenamiento de la información como la provisión de mecanismos para la manipulación de la información, además, un sistema de base de datos debe de tener implementados mecanismos de seguridad que garanticen la integridad de la información, a pesar de caídas del sistema o intentos de accesos no autorizados.

Un objetivo principal de un sistema de base de datos es proporcionar a los usuarios finales una visión abstracta de los datos, esto se logra escondiendo ciertos detalles de como se almacenan y mantienen los datos.

1.2.8.1. Objetivos de un Sistema de Bases de Datos

Los objetivos principales de un sistema de base de datos es disminuir los siguientes aspectos:

- Redundancia e inconsistencia en los datos.

  Información repetida que aumenta el costo de almacenamiento y acceso a los datos. Falta de concordancia entre datos que se supone son iguales. Bajas vs pedidos, padres vs hijos.

- Dificultad para tener acceso a los datos.

  Cubrir las necesidades de información del usuario o entidad, esto implica prevenir cualquier consulta o situación posible de ser solicitada.

- Aislamiento de los datos.

  En las primeras bases de datos se utilizaban grupos de archivos que muchas veces eran de distinto tipo. Hoy en día aún sigue este problema por causa de los malos diseños de bases de datos.

- Anomalías del acceso concurrente.

  Evitar inconsistencias por actualizaciones de usuarios que acceden al mismo tiempo a la base de datos.

- Problemas de seguridad.

La información que se guarda en una base de datos no debe ser vista con la misma profundidad por todos los usuarios. Existen niveles de usuarios y restricciones para consultar la información.

- Problemas de integridad.

  Los datos que ingresan a una base deben estar bien filtrados de manera que no se almacene información errónea o sin el formato adecuado. Para esto se implementan restricciones de integridad basadas en reglas de negocios.

1.2.9. Sistema Manejador de Bases de Datos

Es un conjunto de elementos de software para cumplir con los objetivos de los Sistemas de Bases de datos; Los SMBD proporcionan los mecanismos necesarios para la administración de los datos a través de una Interfaz de Usuario.

Ejemplos de SMBD son: dBase, Fox Pro, IBM db2, Informix, SQL Server, Oracle, PostgreSQL, MySQL, etc.

1.3. Modelos de Bases de Datos

Cuando nos referimos a un Modelo de Bases de Datos es debido a que los datos a almacenar deben mantener una estructura y organización para que sean correctamente almacenados y recuperados, un modelo por sí solo es una abstracción de la realidad que puede ser manipulable con el objeto de buscar una solución a un problema real; en las bases de datos, el problema es almacenar y recuperar datos de forma eficiente para presentar información útil para la toma de decisiones, así pues este problema ha sido solucionado de diversas maneras, ajustando los datos a diversos modelos.

1.3.1. Organización Física de los Datos

Antes de adentrarnos en los modelos de bases de datos, es pertinente revisar cuáles son los medios de almacenamiento utilizados por una computadora y en los cuales se van a organizar físicamente los datos, (Catherine 2004: 761-789)

1.3.1.1. Medios de Almacenamiento

Distintos tipos de medios son utilizados para almacenar, respaldar y procesar bancos de datos.

1. Almacenamiento en disco.

   Las bases de datos son usualmente almacenadas en discos. A diferencia de la memoria principal, el almacenamiento es disco no es volátil, esto es, que no hay pérdida de datos cuando el sistema es apagado. Un disco es un dispositivo de almacenamiento de acceso directo, DASD por sus siglas en inglés (Direct Access Storage Device), cuyo significado refiere a que el acceso a los datos puede ser en diferente orden. Existen diversos formatos de disco, discos duros, discos flexibles, por ejemplo. Usualmente, las bases de datos son almacenadas en discos duros, estas unidades de disco son parte del hardware de un equipo de cómputo, Asimismo las unidades de disco pueden ser portables. Aunque los discos duros son el medio más común para almacenar las bases de datos, estos poseen desventajas causadas por la misma tecnología. Mientras que los datos no son afectados por una falla en los sistemas, el mayor problema ocurre cuando una unidad de disco falla, destruyendo los datos que contenía. Los discos duros están hechos de medios magnéticos que almacenan la información, a su vez cuentan con cabezas para la lectura

y escritura de los datos, las cuales pueden fallar en algún momento, impidiendo el acceso a los datos por fallas en las cabezas lectoras o en algún caso existiendo error en la escritura de datos haciendo imposible su recuperación.

2. Cintas Magnéticas.

   Las cintas magnéticas son medios no volátiles de almacenamiento que proveen acceso secuencial de los datos, esto significa que para encontrar un dato en particular es necesario leer los datos que lo preceden en la cinta magnética. Las cintas no son utilizadas para el procesamiento ordinario de bases de datos que usualmente requieren de acceso directo a los datos. Sin embargo, la cintas magnéticas son utilizadas ampliamente para la realización de respaldo de datos, también es un medio comúnmente utilizado para transferir información entre organizaciones.

3. Memoria Principal.

   La memoria principal es un medio de almacenamiento volátil, es decir que la información que contiene se pierde al reiniciar o apagar el sistema. Provee acceso directo a los datos almacenados. Las bases de datos normalmente no trabajan sobre este medio de almacenamiento debido a que el espacio es limitado y generalmente compartido con otras aplicaciones, el mismo sistema operativo. Su utilización en las bases de datos es restringida a la utilización del *buffer* que es una porción de memoria utilizada para mantener registros en memoria para su procesamiento.

1.3.2. Modelo de Archivos

La organización de los datos en archivos refiere a la forma en que los datos son almacenados y la forma en que éstos son recuperados en el futuro. En el modelo de archivos la estructura básica persistente es el archivo, constituido por un conjunto de campos y registros. Este modelo de bases de datos tiene las siguientes características.

Los datos a almacenar en los archivos deben contar con un formato en particular, explicando consistentemente este punto debemos primero comprender que los sistemas de información que utilizan bases de datos bajo el modelo de archivos utilizan estructuras de datos para el manejo de éstos, tanto para su recuperación, almacenamiento y procesamiento en memoria; es decir los datos captados por el sistema de información se deben alojar en estructuras de datos que son un conjunto de variables interrelacionadas que tienen como objetivo el manejo de datos en memoria.

```
struct estudiantes
{
        char cuenta[9];
        char nombre[35];
        int edad;
        char carrera[10];
        char semestre[2];
};
```

Figura 1. Definición de una estructura de datos en c++.

Los datos manipulados en memoria a través de estructuras de datos son almacenados en archivos. *"Un archivo es un conjunto de datos estructurados en una colección de entidades elementales o básicas denominadas registros o artículos, que son de igual tipo y constan a su vez de diferentes entidades de nivel más bajo denominados campos."* (Joyanes y Zahonero 1998: 584)

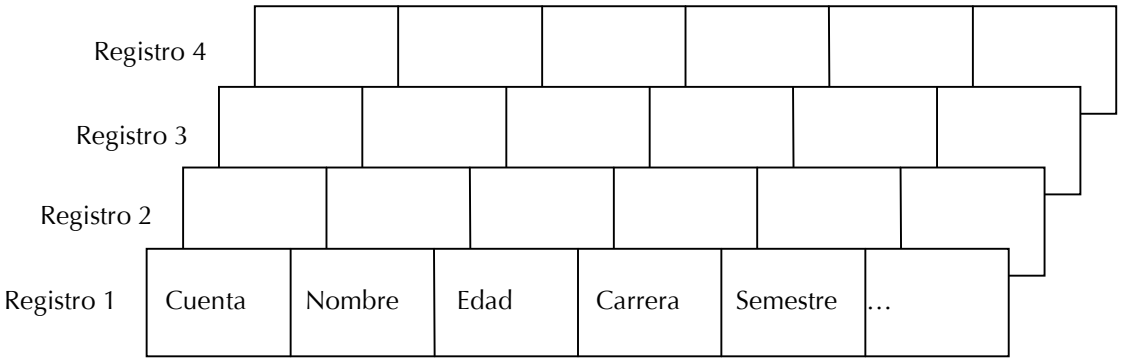

Figura 2. Estructura de un archivo <<Alumnos.dat>>

Bajo el modelo de archivos, se comprende a una base de datos como *"Una colección de archivos a los que puede accederse por un conjunto de programas y que contienen todos ellos datos relacionados constituye una base de datos. Así, una base de datos de una universidad puede contener archivos de estudiantes, archivos de nóminas, inventarios de equipos, profesores, etc."* (Joyanes y Zahonero 1998: 584)

Este modelo de bases de datos también contempla los métodos de acceso a los archivos. Según las características del soporte empleado y el modo en que se han organizado los registros, se consideran dos tipos de acceso a los registros de un archivo:

- Acceso secuencial.

  Implica el acceso a un archivo según el orden de almacenamiento de sus registros, uno tras otro, (Ver figura 3).

- Acceso secuencial indexado.

  Para este tipo de acceso debemos asignar un índice sobre los registros, éste puede ser mediante la elección de un campo sobre el cuál realicemos búsquedas frecuentes, entonces, la información contenida en ese campo para todos los registros deberá colocarse en un archivo por separado (archivo indexado), para la realización de búsquedas en el archivo maestro se consultará primero el archivo indexado que por contar con menos información, realizará una búsqueda más eficiente y así obtendremos la posición del registro que buscamos en el archivo maestro para su posterior apertura.

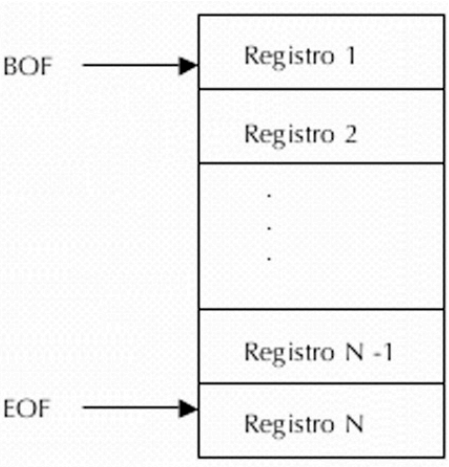

Figura 3. Acceso secuencial

Tras la decisión del tipo de organización que ha de tener el archivo y los métodos de acceso que se van a aplicar para su manipulación, es preciso considerar todas las posibles operaciones que conciernen a los registros de un archivo. Las distintas operaciones que su pueden realizar son:

- Creación.

  Es la primera operación que sufrirá el archivo de datos. Implica la elección de un entorno descriptivo que permita un ágil, rápido y eficaz tratamiento del archivo. Para utilizar un archivo este tiene que existir, la creación exige organización, estructura, localización, o reserva de espacio en el soporte de almacenamiento.

- Consulta.

  Es la operación que permite al usuario acceder al archivo de datos para conocer el contenido de uno, varios o todos los registros.

- Actualización (altas, bajas y modificaciones).

  Es la operación que permite mantener al día el archivo de datos, de tal modo que sea posible consultar el contenido de los registros, agregar nuevos registros, eliminar o actualizar registros.

- Clasificación u Ordenamiento.

  Es una operación muy importante en el modelo de archivos, ya que se realizará una ordenación del contenido del archivo de acuerdo con el valor de un campo en específico, pudiendo ser ascendente ó descendente, alfabética o numérica. (Joyanes y Zahonero 1998: 588)

  La administración de todas estas tareas en el modelo de bases de datos de archivos está a cargo de los Sistemas Manejadores de Archivos, los cuáles se encargarán de cumplir con los objetivos de los Sistemas de Bases de Datos, descritos en el tema *1.2.8 Sistema de Bases de Datos*.

  Una vez comprendida la estructura del modelo de archivos es importante identificar las ventajas y desventajas con las que cuenta.

- Ventajas:
  - Debido a que su aparición fue en los años 50's, en su momento significó un gran avance y permitía realizar un manejo adecuado de la información.

- Desventajas:
  - Debido a que sólo existían dos métodos de acceso, principalmente, secuencial e indexado, el tiempo que tomaba recuperar la información era tardado.
  - El utilizar una gran cantidad de archivos permitía errores de redundancia (duplicidad innecesaria de los datos).
  - El acceso no esta garantizado para todos los usuarios, no soportaba una adecuada concurrencia.
  - No existía seguridad en los datos, ya que cualquier persona puede modificar los archivos sin la necesidad de hacer uso de una interfaz de usuario u otra aplicación específica.

1.3.2.1. Modelo Jerárquico

A principios de los años 60 apareció el modelo jerárquico de bases de datos, los principios que maneja este modelo son basados en el modelo de archivos, ya que la información se sigue almacenando en registros, el cambio primordial existente es la utilización de una estructura de árbol para el manejo de los datos.

En este modelo de base de datos se implementan las relaciones padres a hijos, es decir relaciones de 1:N (uno a muchos) p. ej. 1 alumno puede inscribirse en muchas materias. La implementación de estas relaciones se facilita dada la estructura de árbol que se maneja.

En este modelo se introduce el concepto de *nodo*, que es una estructura persistente dentro de este modelo, un nodo padre puede tener varios nodos hijos y el nodo que no tiene padres, es denominado nodo raíz.

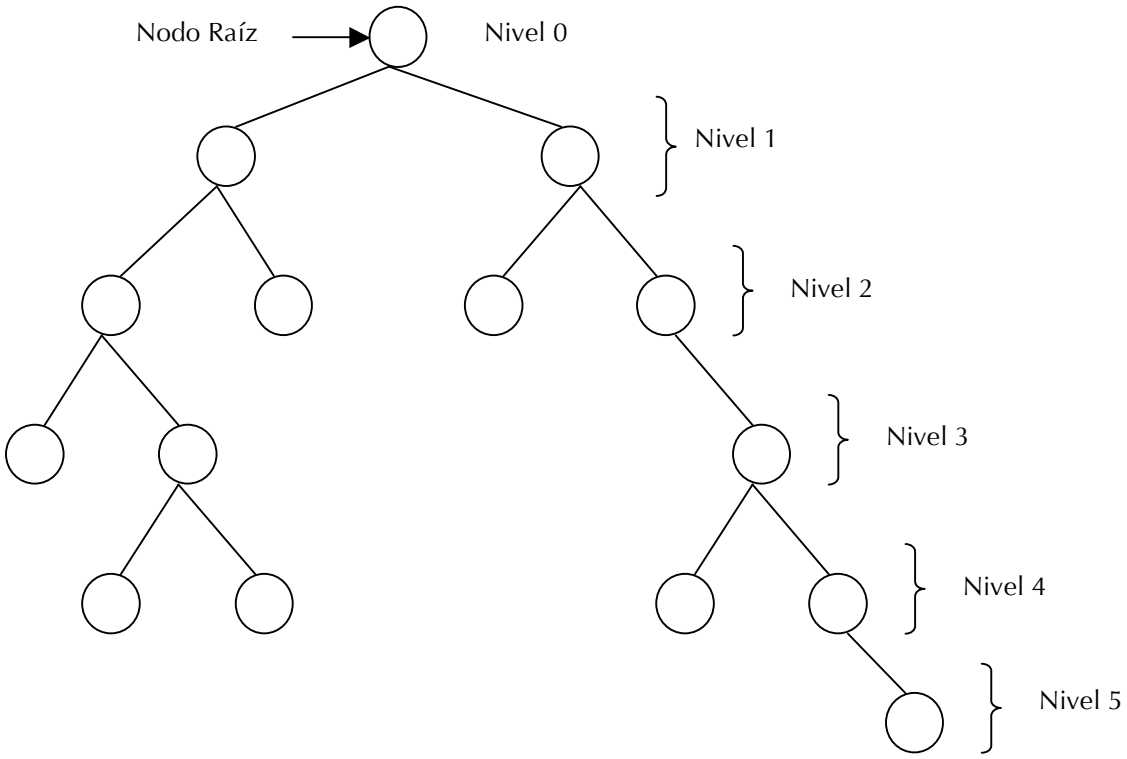

Figura 4. Estructura jerárquica de bases de datos.

Para el manejo de los datos, los Sistemas Manejadores de Bases de Datos jerárquicas utilizan estructuras de datos de árbol, en dónde cada nodo además de contener los datos a almacenar, debe contener la dirección de memoria en dónde se encuentra el nodo padre y su nodo hijo en caso de que éste tenga uno.

La estructura jerárquica tiene como principal desventaja la dificultad para la recuperación a los datos, ya que la navegación en una estructura de árbol requiere de algoritmos complejos.

Asimismo, las desventajas de utilizar un modelo jerárquico es que a pesar de manejar relaciones 1:N no se satisface la regla de integridad referencial, es decir que un nodo hijo debe tener un nodo padre válido. La duplicidad de los registros almacenado también supone una de las desventajas de este modelo.

1.3.2.2. Modelo de Red

El modelo de red surge de igual forma en los años 60, su implementación se debe a que el modelo jerárquico no soporta relaciones N:N y con el objeto de solucionar este problema, es como nace el modelo de red de bases de datos. El modelo de red entonces, es una extensión del modelo jerárquico, en dónde el concepto de nodo se extiende para permitir que un mismo nodo tenga diferentes nodos padre.

Este modelo de red, viene a solucionar problemas de redundancia existentes bajo el modelo jerárquico, sin embargo los algoritmos utilizados para administrar los datos bajo este modelo son en su mayoría complejos.

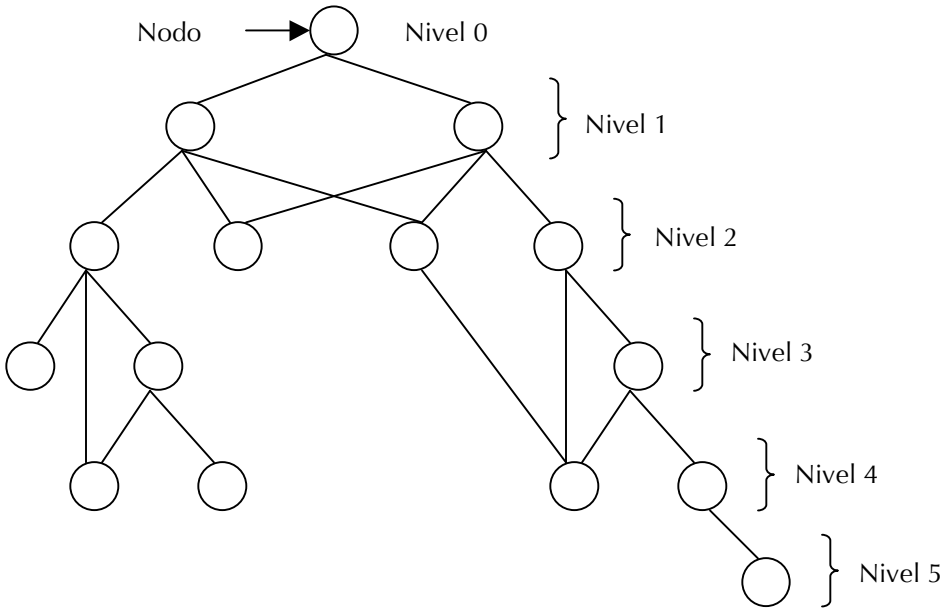

Figura 5. Modelo de red de bases de datos.

1.3.3. Modelo Relacional

El modelo Relacional fue propuesto por Edgar Codd en 1970, en el articulo "A Relational Model of Data For Large Shared Banks". El trabajo de investigación de éste modelo fue realizado por Codd en asociación con IBM en los laboratorios de IBM en San José, California, EUA. Hasta antes de éste modelo, el mercado era dominado principalmente por los sistemas manejadores de bases de datos del modelo de red y jerárquico, los cuáles utilizaban estructuras complejas para almacenar los datos, mismo que hacia difícil la comprensión por parte de los usuarios.

El modelo relacional fue pensado para proveer una excelente fuente de información con la creación de los Sistemas Manejadores de Bases de Datos Relacionales debido a la implementación de técnicas que permiten control sobre la concurrencia, optimización en las consultas, integridad de los datos, administración de transacciones etc. (Catherine 2004:169-170)

El modelo relacional, es un diseño matemático basado en relaciones que tiene como fundamento la Teoría de Conjuntos y Lógica de Predicados; entonces, la estructura persistente de éste modelo es la Relación, que es un conjunto de tuplas y atributos que se componen de un cuerpo y un encabezado.

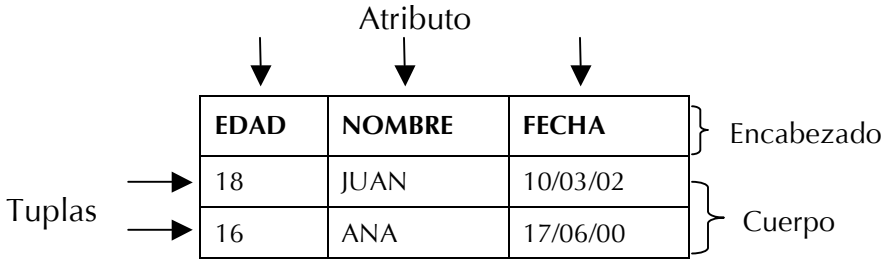

Figura 6. Relación: Estructura persistente del Modelo Relacional.

Adentrándonos en el modelo relacional definimos una tupla como una ocurrencia dentro de una relación y un atributo como una propiedad de la relación.

A su vez el encabezado de una relación es definido como el conjunto de elementos formado por nombres de atributo y tipo de atributo.

P. ej. Encabezado = {(Edad, entero), (Nombre, carácter variable), (Fecha, Fecha)}

El Cuerpo de una relación como un conjunto de pares formado por nombre de atributo y valor.

P. ej. Cuerpo = {(Edad, 18), (Nombre, 'Juan'), (Fecha, '10/03/02'),(Edad, 16), (Nombre, 'Ana'), (Fecha, '17/06/00')}

Edgar Codd también define un conjunto de propiedades para una relación, las cuales son:

- Atomicidad: para cada atributo corresponde un valor
- No puede haber tuplas duplicadas
- Las tuplas no están ordenadas
- Los atributos no están ordenados

Cuando Codd propone el modelo relacional, define un conjunto de 12 regalas que deben cumplir las bases de datos relacionales, también deberán cumplir con un conjunto de formas normales que satisfagan el proceso de normalización descrito por Codd años más tarde.

Reglas de Codd:

- Regla1

  Toda la información dentro de la BD relacional, debe ser presentada en tablas, incluso en Diccionario el Datos.

- Regla2

  Se van a recuperar datos precisos a través de una combinación de columna, tabla y llave primaria (PK).

- Regla3

  Debe manejar valores nulos.

- Regla4

  El diccionario de datos debe tener la misma estructura, estar contenido en tablas, y se va a usar de igual forma que las tablas de usuario.

- Regla5

  Un solo lenguaje es usado para comunicarse con el SMBD y el usuario.

- Regla6

  Vistas actualizables. Una vista es una representación de las tablas dirigida a usuarios específicos.

- Regla7

  Usa algebra relacional.

- Regla8

  Independencia física de los datos.

- Regla9

  Independencia lógica de los datos.

- Regla10

  Reglas de integridad.

- Regla11

  Independencia de distribución.

- Regla12

  Ningún lenguaje diferente a SQL puede cambiar lo que SQL ya definió.

El modelo relacional por tanto consiste en un conjunto de relaciones que se encuentran vinculadas o relacionadas a partir de los atributos que componen a una relación, esto se logra a través de un conjunto de expresiones matemáticas que fundamentan a éste modelo.

Las ventajas que posee el modelo de base de datos relacional es que a través de un buen diseño se puede cumplir con las reglas de integridad de las bases de datos, también es posible el control de la redundancia, control de acceso, y manejo de la integridad de los datos.

Un modelo de bases de datos como se había mencionado con anterioridad constituye un conjunto de elementos y reglas bajo los cuáles se va realizar la organización de los datos. En el modelo relacional se utiliza la relación como estructura persistente, que físicamente posee la representación de una tabla.

EDAD	NOMBRE	FECHA
18	JUAN	10/03/02
16	ANA	17/06/00

Figura 7. Tabla. Representación física de una relación.

Un modelo relacional de bases de datos puede expresarse a través de un conjunto de diagramas, como puede ser un diagrama de entidades ó un diagrama Entidad – Relación; el objetivo de éste tipo de diagramas es mostrar las forma en que se va a constituir una base de datos relacional.

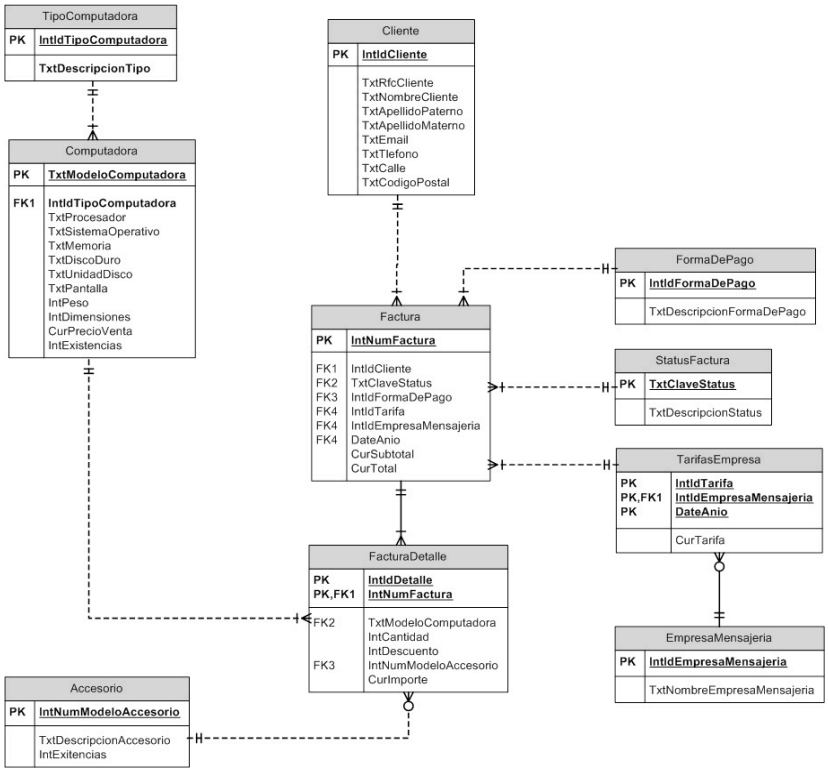

Figura 8. Diagrama Entidad – Relación, ilustrativo de un modelo de base de datos relacional

Un elemento más de éste modelo relacional es la inclusión de un lenguaje formal que será utilizado por los SMBD relacionales para la creación, modificación y eliminación de objetos de la base de datos así como para la inserción, actualización, y eliminación de registros y para la recuperación de información de la base de datos relacional. El lenguaje que utilizarán estos SMBDR es el Lenguaje Estructurado de Consultas SQL.

Los SMBDR por tanto deberán cumplir con los objetivos de todo SMBD y a su vez cumplir con las reglas propias del modelo relacional.

En la actualidad son los sistemas de bases de datos bajo el modelo relacional los que predominan en el mercado ya que poseen la virtud de realizar un manejo dinámico de la información además de que para el usuario final no tiene relevancia la forma en que se almacenan los datos, es decir los SMBDR poseen una independencia física y lógica de los datos. Aunado al modelo relacional esta definido el Lenguaje Estructurado de Consultas, el cual permite a través de un conjunto de operaciones la Definición de los objetos de la base de datos, la Manipulación y el Control de acceso a los mismos. El SQL a través de su operador más poderoso permite la recuperación de la información de forma precisa.

Indudablemente realizar el planteamiento del modelo relacional en unos cuantos párrafos es imposible, la intensión es presentar al lector un esbozo del funcionamiento de los diversos modelos de bases de datos con el objeto de que reflexione sobre las ventajas y desventajas de cada uno de ellos, esto a su vez le permitirá al lector obtener un panorama más amplio sobre las necesidades actuales en cuanto al almacenamiento de la información y con ello adentrarnos en el tópico central de ésta tesis: La minería de datos.

1.3.4. Modelo Orientado a Objetos

Éste modelo de bases de datos es relativamente reciente, su aparición se debe a la expansión de los sistemas de información desarrollados bajo el paradigma orientado a objetos y surge con la necesidad de implementar bases de datos que soporten elementos de dicho paradigma como lo son la herencia, el encapsulamiento y polimorfismo.

La estructura persistente de este modelo es el objeto, constituido por atributos y métodos.

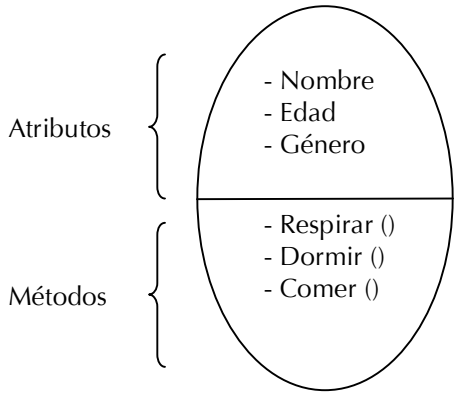

Figura 9. Objeto, estructura persistente del modelo de base de datos orientadas a objetos.

El objetivo de este modelo de bases de datos es realizar una transición transparente de los datos desde el sistema de información hacia la base de datos orientada a objetos. Debido a que este modelo de base de datos es reciente no existe una unificación respecto a la definición formal del modelo; una de las teorías supone que para conceptualizar un modelo de base de datos orientado a objetos debe realizarse un diagrama de clases propio de una metodología de desarrollo de sistemas orientados a objetos.

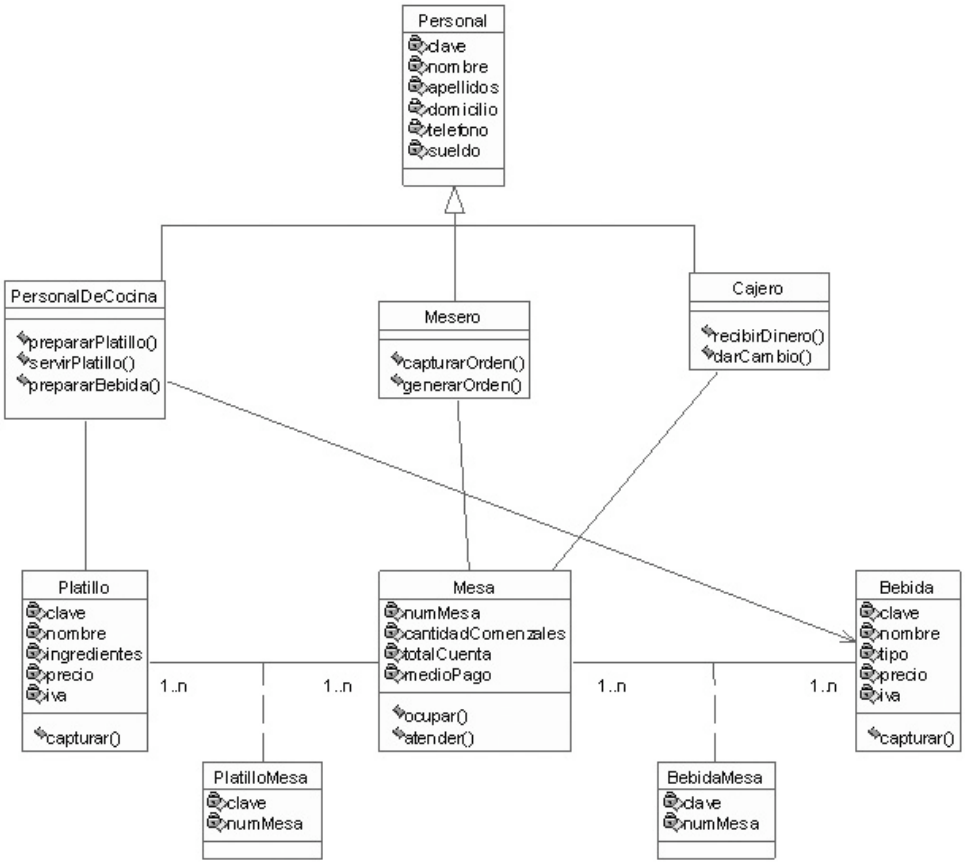

Figura 10. Diagrama de clases.

1.4. Necesidades actuales

En la actualidad, las organizaciones utilizan principalmente los Sistemas Manejadores de Bases de Datos Relacionales ya que las características de éstos les permiten una adecuada administración de la información; sin embargo, cuando se desea manejar un volumen de datos extraordinario es necesario mirar hacia otros horizontes.

Hoy día existen grandes organizaciones que tienen una presencia a nivel nacional o multinacional y llevar el manejo de su información en una sola base de datos resulta imposible, ya que los volúmenes de información que maneja son realmente grandes, para éstas situaciones el modelo relacional resulta obsoleto y por ello se requirió en su momento de la generación de nuevas tecnologías, tal es el caso de las Bases de Datos Distribuidas, las Bases de datos Multidimensionales, los Cubos de Información, etc.

Al contar con volúmenes extraordinarios de información la recuperación de datos específicos se vuelve un problema importante ya que las organizaciones requieren conocer éstos de forma precisa y en tiempo real con el objeto de tomar decisiones que encaminen a la organización por un camino promisorio, por tanto podemos afirmar que la presentación de la información en el momento justo para la toma de decisiones se ha vuelto un factor crítico de éxito en las organizaciones.

La cantidad de información colectada y presentada por los sistemas de información de las organizaciones hoy en día tiene como repercusión grandes volúmenes de datos que son apreciados de forma particular, haciéndose difícil de apreciar en algunos casos, de forma conjunta o en grupos más sólidos el fenómeno bajo

el cual se desenvuelve el Modelo de Negocio de las Organizaciones; es decir que a mayor volumen de datos se tiene una percepción menos detallada del entorno bajo el cual se desenvuelven las organizaciones.

Las necesidades actuales de las organizaciones, entonces, requieren no sólo de tecnologías específicas para almacenar la información sino también de un conjunto de herramientas que permitan obtener un conocimiento sustancial del modelo de negocio bajo el cual se desenvuelven, por ello han surgido técnicas como la Minería de Datos que van a permitir la extracción de nueva información que se halla oculta en las bases de datos actuales con el objeto de generar una ventaja competitiva para la evolución de las organizaciones. La Minería de Datos, entonces, intenta ir un paso más allá de la información con la cual se cuenta, la Minería de Datos generará conocimiento a partir de la historia y del entorno de las organizaciones con el objeto de presentar nuevos elementos que intervengan en el proceso de toma de decisiones.

Así pues, el hombre de hoy día tiene grandes necesidades respecto al manejo y administración de su información, no sólo conservándola y presentándola en el momento justo, sino que también requiere analizarla de forma particular para retroalimentarse y con ello obtener nuevo conocimiento que le permita darse cuenta del entorno bajo el cual se desenvuelve.

Este documento de tesis por tanto no podría abarcar todas las posibilidades respecto a las necesidades actuales del hombre en cuanto al manejo de la información, el enfoque que se hará a partir del capítulo siguiente estará orientado hacia la extracción de conocimiento en las bases de datos a través del conjunto de técnicas que maneja la Minería de Datos.

1.5. ¿Qué es la Minería de Datos?

La Minería de Datos en principio podemos definirla como un conjunto de técnicas matemáticas, estadísticas y computacionales que van a permitir la extracción de conocimiento de las bases de datos.

La mayor razón por la cuál la Minería de Datos ha llamado la atención de las organizaciones en los últimos años es debido a que ésta tiene la gran habilidad de ser utilizada en grandes volúmenes de datos para trasformarlos en información útil y conocimiento. (Han y Micheline 2001:1-5)

La Minería de Datos puede ser vista como el resultado natural de la evolución de las tecnologías de la información de las bases de datos, dentro del área del análisis de datos.

La Minería de Datos por tanto, nace de la necesidad de las organizaciones por conocer patrones o características ocultas dentro de la información que manejan, por ello la extracción de conocimiento se consolidó dentro de un área conocida como Minería de Datos.

Una vez que se tienen presentes las necesidades históricas del hombre por preservar su información, la manera en que históricamente la ha almacenado y las necesidades actuales antes los grandes volúmenes de información, es como podemos adentrarnos en el tópico central de ésta tesis, cuyo enfoque esta pensado hacia las técnicas de minería de datos que van a permitir la generación de conocimiento ante grandes volúmenes de información.

Capítulo 2

Minería de Datos

Una vez que las organizaciones implementan un sistema de Bases de Datos, es predecible que en un par de años después (dependiendo la carga transaccional), cuenten con una gran repositorio de datos, con un volumen de información importante; cuando esto sucede los datos almacenados se dejan de ver de forma particular, es decir, la información que es extraída de la Base de Datos se utiliza para propósitos cada vez más generales, dejando de lado una inspección minuciosa de los datos almacenados.

Los datos almacenados de forma cuantiosa en una Base de Datos tienen la potencialidad de explicar fenómenos del entorno bajo el cuál se desenvuelve el modelo de negocio de una organización, y quizá ¿por qué no? ir un paso más allá para predecir el comportamiento de los datos.

2.1. Definición

La Minería de Datos es el proceso por el cuál atraviesa un conjunto de datos almacenados, con el objeto de hacer visibles patrones ocultos que proporcionen conocimiento entendible y útil para el dueño de los datos, a través de técnicas matemáticas, estadísticas y computacionales.

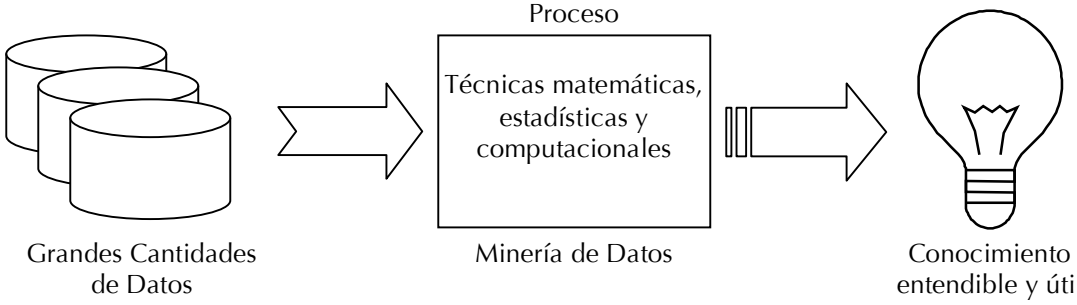

Figura 11. Minería de Datos.

Otras definiciones de Minería de Datos son:

> *"Minería de Datos es la exploración y el análisis de grandes cantidades de datos con el objeto de descubrir patrones y reglas significativas."* (Berry y Linoff 2004: 7)
>
> *"La Minería de Datos es descrita como el proceso de describir patrones en datos. El proceso debe ser automático o (más usualmente) semiautomático. Los patrones descubiertos deben ser*

> *significativos desde el punto de vista en que deben otorgar alguna ventaja, usualmente económica. Los datos invariablemente deben presentarse en grandes cantidades."* (Witten y Eibe 2005: 5)
>
> *"La Minería de Datos es un campo interdisciplinario en donde interactúan técnicas como máquinas de aprendizaje, reconocimiento de patrones, la estadística, bases de datos, y visualización de direcciones para la extracción de información en grandes bancos de datos."* (Larose 2005: 2)

Haciendo referencia a la primera definición hecha en este capítulo, podemos afirmar que la Minería de Datos es un proceso por que existe un conjunto de tareas que transforman los datos a través de técnicas matemáticas, estadísticas y computacionales, entregando como salida conocimiento entendible y útil. ¿En que consisten dichas tareas?, o mejor dicho ¿en qué consiste el proceso de Minería de Datos y cuáles son las técnicas utilizadas? Sin duda, no es una respuesta sencilla, mucho depende de los datos a analizar, de la perspicacia y experiencia de quien analiza los datos, pero sobre todo depende de lo que se quiera: *describir* o *predecir* un fenómeno.

Por su puesto, la Minería de Datos puede entenderse desde dos perspectivas, dando lugar a dos modelos de Minería de Datos:

- Modelo Deductivo.
- Modelo Predictivo.

2.2. Modelos de Minería de Datos

2.2.1. Modelo Deductivo

El modelo deductivo de Minería de Datos tiene por objeto realizar el análisis de un fenómeno cuyo tiempo y espacio son fijos, suponga por ejemplo que se ha tomado una panorámica en el tiempo y se desea explicar lo que allí ocurrió, por tanto, el fenómeno a explicar tiene la característica de que carece de historia, no se conoce que ocurrió antes, es más puede que no tenga historia el fenómeno a explicar; un ejemplo de un fenómeno en dónde se utilizaría el modelo deductivo de Minería de Datos es el siguiente:

Aplicación del modelo deductivo de Minería de Datos en el campo médico.

En medicina, el término cáncer es usado para identificar una entidad clínica y anatomopatológica de carácter maligno que afecta a un paciente, y cuyas características histopatológicas son la alteración morfológica y funcional seguida de la proliferación descontrolada —no siempre acelerada— de las células de un tejido que invaden, desplazan y destruyen, localmente y a distancia, otros tejidos sanos del organismo. (Web. 02)

En la lucha por la detección a tiempo de células dañinas y con el objeto de evitar el cáncer, médicos expertos realizan análisis sobre las células de los pacientes, el análisis consiste en evaluar las características de las células como son tamaño y forma, el médico experto realiza un conteo de las zonas anómalas y con ello determina un diagnóstico clínico sobre células dañadas. El médico experto requiere de un arduo entrenamiento para identificar claramente cuáles son las zonas dañadas de célula, a demás de contar con cierta habilidad en su trabajo.

Una aplicación de Minería de Datos bajo el modelo deductivo sería la generación de una red neuronal artificial que formaría parte de un sistema experto en la detección de células dañinas en el organismo de una persona; la red neuronal en cuestión se avocaría a la identificación de zonas anómalas en una célula, esto sería posible a través del entrenamiento de la red neuronal tomando como base la experiencia de un médico experto.

En este caso de estudio no existe una historia como tal, sino que se desea dar explicación a un fenómeno que puede detenerse en el tiempo para su estudio. El modelo deductivo de Minería de Datos se avoca a dar explicación a fenómenos que están retratados como una panorámica en el tiempo, y cabe resaltar que los fenómenos son estudiados de forma universal, primeramente, para dar una explicación sobre casos ya muy particulares, después.

2.2.2. Modelo Predictivo

El modelo predictivo de Minería de Datos tiene la característica principal de que la explicación que se da de los fenómenos estudiados es la predicción de un evento futuro. Esto es posible debido a que el fenómeno posee la propiedad histórica suficiente para poder realizar tareas estadísticas, matemáticas y computacionales que arrojen una predicción sobre eventos futuros dentro de este fenómeno. Un ejemplo sobre este modelo de Minería de Datos se presenta a continuación:

Aplicación del modelo predictivo de Minería de Datos en el ámbito empresarial; marketing y ventas.

Las organizaciones empresariales cuentan con información basta de las operaciones que realizan diariamente con motivo de las ventas. Los clientes, como siempre, son el objetivo principal de este tipo de estudios en éste campo de aplicación debido a que se persiguen fines económicos tales como: mantener un cliente, ofrecer con la seguridad de compra un producto a un cliente específico, incrementar las ventas, etc.

Supongamos una base de datos de clientes en dónde se tiene la posibilidad de conocer el histórico de compra de los diferentes clientes del negocio, entonces, esta historia generada puede ser utilizada para discriminar y hacer evidentes los patrones de conducta de compra de un cliente o conjunto de clientes. Esto es posible a través de un análisis de *clusters*, en dónde se realizan agrupaciones dado un espacio muestral con el objeto de delimitar aquellos clientes que son *"más cercanos"*, es decir que tienen características similares en su comportamiento, en sus patrones de consumo.

Una vez encontrados los grupos ó *clusters* que existen en nuestros clientes, es posible realizar una predicción sobre los siguientes productos que serán adquiridos por un cliente o, por que no, dar un paso adelante ofreciendo un producto determinado a un cliente, también determinado, con la seguridad de que será adquirido por éste.

En este modelo de Minería de Datos, por tanto, es de suma importancia contar con registros históricos del fenómeno a estudiar ya que serán vitales estos sesgos en el tiempo para dar una explicación y predicción del fenómeno en cuestión.

2.3. Funcionalidades de la Minería de Datos

Las funcionalidades de la Minería de Datos son descritas como un conjunto de tareas que pueden ser desarrolladas en los datos a examinar. Una funcionalidad o tipo de estudio de Minería de Datos puede ser válida dentro del modelo deductivo o predictivo; ¿cuál es la funcionalidad más apropiada para un conjunto de datos? Depende del tipo de datos con que se cuente, del objetivo del estudio, del modelo de Minería de Datos bajo el cuál se vaya a trabajar, pero sobre todo de la experiencia del experto en Minería de Datos, que con su criterio seleccionará de un conjunto de funcionalidades cual es la más apropiada.

Las tareas que se llevan a cabo más comúnmente en Minería de Datos son:

- Descripción.
- Estimación.
- Clasificación.
- Agrupamiento.
- Predicción.
- Asociación.

2.3.1. Descripción de Clases/Conceptos: Caracterización y Discriminación

Generalmente los datos pueden estar agrupados en clases o conceptos. Por ejemplo, en una tienda de aparatos electrónicos, los productos pueden estar agrupados en clases como computadoras, impresoras, consumibles, etc., y a su vez los clientes pueden estar también agrupados a través de conceptos generados por los lineamientos de la tienda, es decir clientes buenos, clientes malos, etc.

Así, una funcionalidad de la Minería de Datos es la descripción de clases/conceptos a través de la sumarización concisa y precisa de diversos elementos. La descripción de clases/conceptos puede realizarse a través de dos vías: la caracterización de los datos y a través de la discriminación de los datos.

2.3.1.1. Caracterización

Es la sumarización de las características o rasgos generales del conjunto de datos a analizar. Los datos corresponden a las especificaciones del propietario de los datos; este tipo de sumarizaciones y agrupaciones pueden ser fácilmente realizadas a través de consultas SQL en la base de datos relacional. Cuando la sumarización de las características de los datos debe estar siempre presente en el modelo de negocio, entonces se puede recurrir a un Data WareHouse en dónde a través de un cubo OLAP se realice la sumarización y caracterización de los datos de forma optimizada.

El resultado de la caracterización de los datos puede ser mostrado a través de: **gráficas circulares ó de pastel, gráficas de barras, cubos multidimensionales de datos, tablas multidimensionales, tablas cruzadas.**

El resultado de la descripción de los datos puede ser presentado en relaciones generalizadas o en **reglas de caracterización**. (Han y Micheline 2001: 21)

- Caso de estudio

  Imagine el proceso de Minería de Datos con análisis Descriptivo a través de la caracterización de los datos de consumidores en una tienda de electrodomésticos, entonces, el objetivo de dicho sistema será sumarizar las características de los consumidores que gastan más de $3,000 pesos en electrodomésticos. El resultado podría ser un perfil generalizado de clientes en dónde los que cumplen con la condición son adultos mayores entre 35 – 40 años, con empleo fijo y un excelente historial crediticio. En este caso de estudio se agrupan todas las características de los clientes que cumplieron con una condición determinada para obtener un perfil general de consumidores.

2.3.1.2. Discriminación

La discriminación de los datos es la comparación de los rasgos generales de los datos con el objeto de agruparlos en clases predefinidas. El conjunto de clases objetivo puede ser especificado por el dueño de los datos; este tipo de discriminaciones pueden realizarse a través de consultas SQL en una base de datos relacional. Los métodos utilizadas en la discriminación son similares a los utilizados en la caracterización, aunque en la discriminación de datos es recomendable incluir un conjunto de medidas comparativas que ayuden a distinguir entre la agrupación de datos realizada. A su vez, la presentación

de los resultados finales del estudio de Minería de Datos a través de la Descripción pueden ser presentados a través de **reglas de discriminación**. (Han y Micheline 2001: 21)

- Caso de estudio

  Imagine el proceso de Minería de Datos con análisis Descriptivo a través de la discriminación de los datos de consumidores en una tienda de electrodomésticos en dónde el objetivo del estudio es realizar una comparación entre dos grupos de consumidores en dónde el grupo A de consumidores realiza compras frecuentes (más de dos veces al mes) y en grupo B de consumidores realiza compras más esporádicas (menos de tres veces al año). El resultado del análisis Descriptivo a través de discriminación arrojaría como resultado un perfil comparativo general de los consumidores en dónde el 80% de los consumidores que frecuentemente hicieron compras son personas entre 20 y 40 años de edad con una educación universitaria, y que el 60% de los consumidores que rara vez compraron productos en el año son personas de edad muy avanzada o muy jóvenes.

2.3.2. Estimación

La estimación en Minería de Datos refiere al proceso de hallar una o más funciones o modelos que permitan distinguir y describir un conjunto de datos. La estimación generalmente se basa en un análisis de regresión para estimar el valor de una variable bajo diversas condiciones conocidas. La estimación se hace con variables numéricas ya que el la estimación interviene un modelo matemático que realiza una regresión de datos para generar funciones que permitan estimar valores futuros. (Berry y Linoff 2004: 9)

- Caso de estudio

  A través de un estudio de estimación se va a determinar el monto que una persona va a gastar en un supermercado tomando como base variables numéricas como cantidad invertida, salario o sueldo percibido, número de artículos comprados, etc. Una vez discriminados los datos se obtiene el historial de todos aquellos consumidores que caen dentro de ese grupo, se grafica el gasto que han tenido históricamente en la tienda y a partir de esos datos se realiza una regresión lineal con la finalidad de construir un sistema de ecuaciones que den explicación al fenómeno.

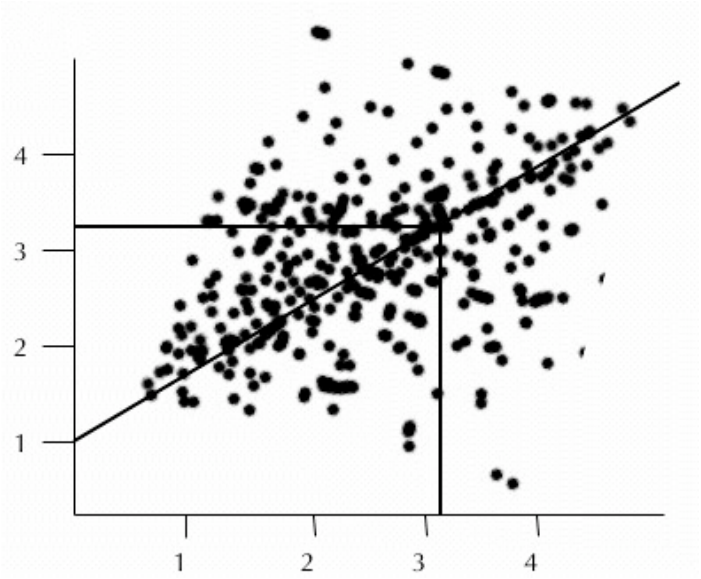

Figura 12. Análisis de regresión.

En la figura anterior, los puntos significan un registro sobre el consumo de un grupo de clientes, la línea que atraviesa diagonalmente la nube de puntos representa el sistema de ecuaciones que dan explicación al fenómeno y a través de este sistema realizar una aproximación sobre el consumo que tendrá un cliente con las características del grupo antes mencionadas.

2.3.3. Clasificación

A diferencia de la estimación, en la clasificación el objeto de estudio son variables categóricas[1]. La clasificación es una de las tareas más comunes que se realizan en Minería de Datos, esto es debido en gran medida a que el hombre siempre ha tenido la necesidad de agrupar y clasificar las cosas presentes en su entorno con el objeto de tener un mayor control de las situaciones.

La clasificación es el proceso de hallar un conjunto de modelos (funciones) que describen y distinguen clases de datos o conceptos, tomando como base variables categóricas; dicho proceso consiste en examinar las características de nuevos registros de datos para asignarlos a un conjunto de *clases predefinidas*. (Han y Micheline 2001: 24)

Así pues, la tarea de clasificación se caracteriza porque previo al estudio Minería de Datos ya se cuenta con la definición de las clases y con un conjunto de datos que ya han sido previamente clasificados, la tarea de clasificación, como ya se ha mencionado, consiste en elaborar un modelo que se aplique a datos que no han sido clasificados, con el objeto de asignarles una categoría o clase predefinida. (Berry y Linoff 2004: 9)

- Caso de estudio

  Suponga que el gerente de una tienda de electrónica desea clasificar el conjunto de productos que ofertan y para ello ha determinado tres categorías principales: a) productos con buena respuesta en el mercado, b) productos con respuesta media en el mercado y c) productos con mala respuesta en el mercado. Por lo tanto, el objetivo del estudio de Minería de datos sería la generación de un modelo que permita clasificar los productos existentes tomando como punto de partida las siguientes variables: precio, marca, ubicación en la tienda y tipo de producto. La representación del conocimiento para este ejemplo se podría realizar a través de un árbol de decisiones que muestre el factor que mejor distinga a cada una de las tres clases.

2.3.4. Agrupamiento (Clustering)

A diferencia de la asociación, en dónde se cuenta con un conjunto de clases predefinidas, en la tarea de agrupamiento se analizan conjuntos de datos con la finalidad de generar clases, este análisis se realiza porque en principio no se conocen las clases que puedan existir en los datos. Para dicho estudio los datos son agrupados bajo el principio de *"dividir un conjunto de objetos en grupos (clusters) de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y los de los objetos de clusters diferentes sean distintos (aislamiento externo del grupo)."*, que refiere a que se deben encontrar los datos con comportamiento similar para agruparlos y a su vez separarlos de un conjunto de datos con características diferentes. (Han y Micheline 2001: 25)

- Caso de estudio

  Continuando con el mismo orden de ideas, en la tienda de electrónica objeto de nuestros casos de estudio, se desea conocer la homogeneidad existente entre los diversos clientes, objetivo del estudio

[1] Variable categórica es aquella que puede adquirir un valor ubicado dentro de un dominio de datos.

de Minería de Datos sería conocer los grupos de clientes con los que se cuenta para implementar diversas estrategias de marketing.

Figura 13. Grafico en D2 representativo del agrupamiento de clientes.

El resultado del análisis de clusters como se puede ver en el gráfico anterior arroja como resultado el agrupamiento de clientes de a cuerdo a la ciudad de dónde son originarios.

2.3.5. Predicción

La tarea de predicción de Minería de Datos es igual a la tarea de clasificación o estimación, excepto que los registros de datos son clasificados con el objeto de realizar predicciones futuras sobre compartimiento ó valor futuro en una variable. Los métodos utilizados en las tareas de estimación y clasificación pueden ser utilizados, bajo circunstancias apropiadas, en la predicción; dichos métodos pueden ser métodos estadísticos tradicionales para puntos de estimación y estimaciones de intervalos, regresión linear simple, regresión múltiple, análisis de auto correlación, redes neuronales artificiales, árboles de decisión, etc. (Larose 2005: 13)

La tarea de predicción es realizada a posteriori de la clasificación, estimación ó agrupamiento.

Realizar la predicción de valores futuros o de comportamiento, requiere de una cantidad de datos históricos importante, dichos datos deben pertenecer al fenómeno que se desea predecir, así pues, se genera un modelo en dónde a través del entrenamiento con los datos históricos es capaz de predecir acontecimientos futuros.

- Caso de estudio

  Suponga que en una planta de producción de juguetes se desea calcular cuál será el nivel de inventario con el que se contará en el trimestre último del año. A través de un análisis predictivo, puede entrenarse un modelo que basándose en los históricos del volumen del almacén se generé la predicción sobre el volumen del trimestre siguiente.

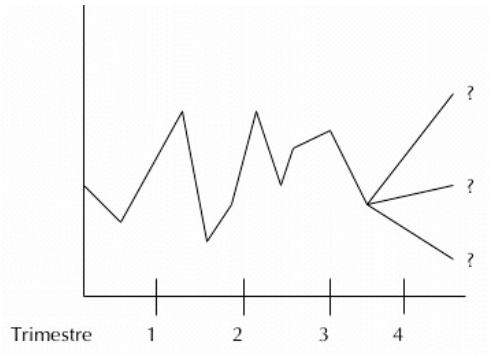

Figura 14. *Predicción del volumen de inventario al último trimestre del año.*

2.3.6. Asociación

La tarea de asociación es aquella que consiste en la determinación de reglas que permitan identificar la relación existente entre valores de atributos y condiciones que frecuentemente vienen juntas. (Han y Micheline 2001: 23) El estudio de asociación permite identificar cuáles son los atributos que siempre coinciden en un fenómeno.

- Caso de estudio

  En una tienda de autoservicio se desea conocer cómo es la venta de sus productos, para ello se ha determinado realizar un análisis de afinidad en dónde se observe cuál es el patrón de venta de sus productos, para ello primero se ha realizado una restricción de los datos en dónde se han seleccionado a todos los clientes que han realizado compras los viernes por la noche. Como resultado del análisis de afinidad se han podido construir reglas de asociación que muestran que los clientes que compran en viernes, siendo varones, adquieren cerveza y pañales; la razón de la compra es simple, los varones son padres de familia y entendiendo que deben pasar el fin de semana en casa cuidando a sus hijos pequeños deciden comprar pañales para no salir de casa y a su vez cerveza para pasar el tiempo.

  Una regla de asociación formalmente puede expresarse de la siguiente manera: día(Y,"Viernes") /\ hora(Y,"Noche") /\ género(X, "masculino") /\ estado civil(X, "casado") /\ hijos(X, "bebé") → compras realizadas (X,"pañales", "cerveza").

  En dónde "X" representa a un consumidor y "Y" un momento.

Capítulo 3

Metodología para Minería de Datos: CRISP – DM 1.0

Como bien ya se ha mencionado, la Minería de Datos es el proceso por el cuál atraviesa un conjunto de datos almacenados con el objeto de hacer visibles patrones ocultos que proporcionen conocimiento entendible y útil para el dueño de los datos, a través de técnicas matemáticas, estadísticas y computacionales.

Así pues, el contexto bajo el cuál se realizará la construcción de un Modelo de Minería de Datos debe ser analizado profundamente con la finalidad de seguir un conjunto de pasos y tareas que mejor satisfagan las exigencias del proyecto y que a su vez se acoplen con el modelo de negocio que sigue la organización, por tanto, utilizar una Metodología para llevar el proceso de Minería de Datos significaría la definición de un proceso de desarrollo, entendiendo por esto último, la forma de trabajo y las guías de acción a seguir en cada una de las etapas del proceso de Minería de Datos.

Hacer uso de una Metodología para llevar a cabo el proceso de Minería de Datos no garantiza el éxito del proyecto ni la calidad del producto final, sin embargo se sientan las bases para el desarrollo y se define un marco de trabajo adecuado, el cuál permita llevar a cabo una mejor gestión en el proceso de Minería de Datos.

Si bien hasta este punto se ha hablado de la Minería de Datos como un proceso y se han revisado cuales son los modelos existentes y las técnicas que son utilizadas, es momento de adentrarnos en dicho proceso y ver cuales son las etapas y tareas que conforman un proceso típico de Minería de Datos, esto lo haremos bajo el enfoque de la Metodología CRISP –DM, la cuál conoceremos en este capítulo y aplicaremos en el caso de estudio objeto de este documento de tesis.

3.1. CRoss – Industry Standard Process of Data Mining

CRISP – DM 1.0, es una metodología para procesos de Minería de Datos, concebida a partir de 1996 y terminada en su primera versión en el año 2002, es desarrollada por un conjunto de "veteranos" especialistas en Minería de Datos, provenientes de 3 grandes compañías como lo son SPSS, NRC y Daimler Crysler, todas ellas enfocadas a la explotación de grandes cantidades de datos.

La experiencia de los fundadores de CRISP, hace posible la versión 1.0 de esta metodología, que en primera instancia podemos describirla como un modelo jerárquico de procesos, consistentes en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo particular): fase, tarea genérica, tarea especializada e instancia de proceso.

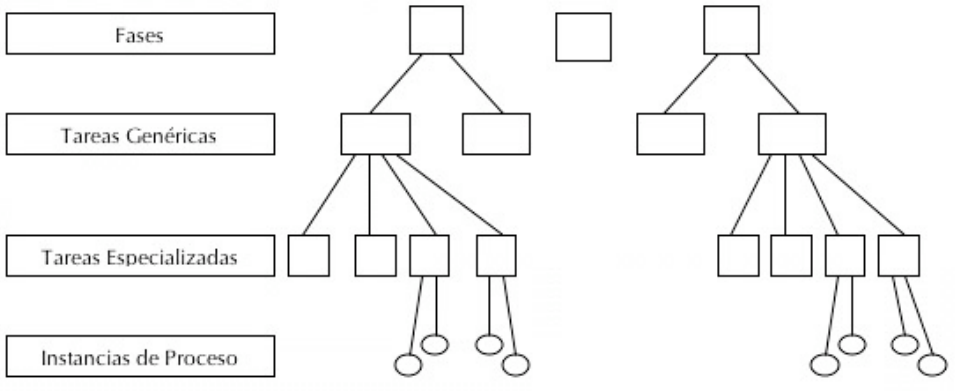

Figura 15. Niveles jerárquicos de CRISP –DM.

El nivel principal del proceso de Minería de Datos, se encuentra organizado en un número de fases, en dónde cada una consiste en un conjunto de tareas genéricas de segundo nivel.

El segundo nivel es llamado genérico por que con él se intentan cubrir todas las posibilidades respecto de las aplicaciones de Minería de Datos.

El tercer nivel, de tareas especializadas, es el lugar en dónde se describe como llevar a cabo las tareas genéricas en situaciones específicas. Por ejemplo, una tarea de segundo nivel, es decir genérica, podría ser el preprocesamiento de los datos, y acorde con ello, una tarea específica de tercer nivel debería estar enfocada al preprocesamiento de datos numéricos, de variables categóricas u otras.

El cuarto nivel del proceso, es el conjunto de acciones, decisiones y resultados, propios de una aplicación de Minería de Datos. Una instancia de proceso es organizada acorde a las tareas definidas en niveles superiores, pero representativa de la aplicación particular a desarrollar.

Si bien CRISP – DM se ha definido como un modelo jerárquico, éste únicamente sirve para conceptualizar los diferentes niveles existentes en la metodología, ya que como ciclo de vida a implementar se utiliza el siguiente modelo de referencia:

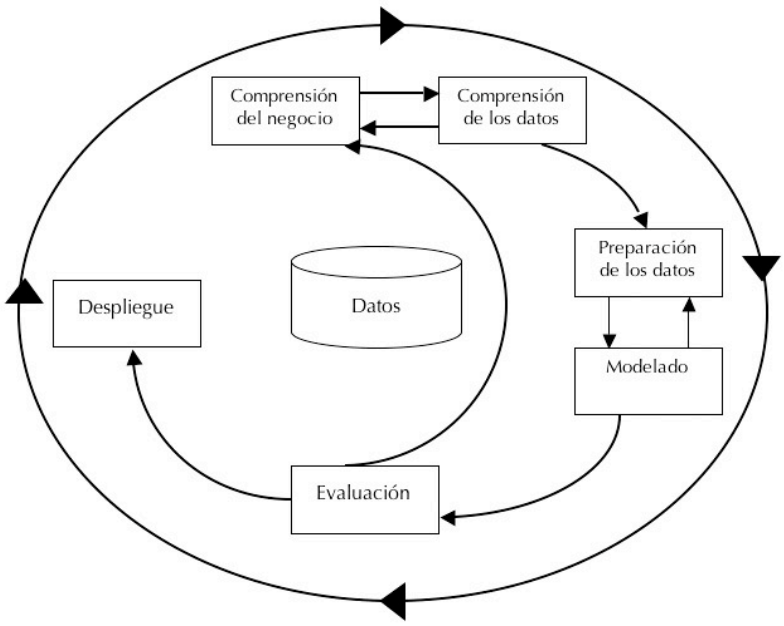

Figura 16. Fases del modelo de referencia CRISP – DM.

El ciclo de vida de un proyecto de Minería de Datos consiste de seis fases; acorde a la figura anterior, podemos decir que la secuencia existente entre las diversas fases no es obligatoria, ya que siempre es necesario avanzar o retroceder entre éstas con el objeto de refinar una tarea que sirva de entrada a una fase siguiente, a su vez, el circulo exterior simboliza el ciclo natural existente en la Minería de Datos, ya que los fenómenos a estudiar son cambiantes a través del tiempo.

A continuación, daremos una breve explicación de cada fase, para más adelante profundizar en ellas.

- Comprensión del negocio.

  Es la fase inicial cuyo propósito principal es lograr la comprensión de los objetivos y requerimientos del proyecto desde la perspectiva particular del negocio, con el fin de traducir el conocimiento adquirido en una definición del problema de Minería de Datos y en el diseño de un plan preliminar.

- Comprensión de los datos.

  Esta fase comienza con una colección inicial de datos en dónde las actividades a realizar van encaminadas a familiarizarse con los datos, evaluar la naturaleza de éstos con el fin de distinguir las variables a utilizar, de identificar las agrupaciones posibles y sobre todo de plantear una hipótesis inicial.

- Preparación de los datos.

  La fase de preparación de datos, consiste en un conjunto de actividades que permita la construcción del conjunto de datos a utilizar durante el estudio; las tareas de ésta fase incluyen la selección de tablas a utilizar, elección de atributos, generación de variables, agrupación de datos y limpieza de los mismos.

- Modelado

  En esta fase se selecciona de varias técnicas la más óptima para el proyecto de Minería de Datos, se generan los algoritmos necesarios y se evalúan constantemente hasta obtener un resultado satisfactorio; entendiendo que pueden existir diversas técnicas de Minería de Datos que dan solución a un mismo problema, hay que evaluar los requerimientos en cuestión de tipos de datos y preparación para su utilización, resulta muy frecuente regresar a la fase de Preparación de datos, para trabajar con una técnica en particular.

- Evaluación

  Durante esta fase el proyecto debe ya contar con un conjunto de modelos, que dan la perspectiva de conocimiento a los datos, el objetivo de esta fase es determinar si el conocimiento generado se encuentra acorde al modelo de negocio y si en verdad refleja nuevo conocimiento, quien es el dueño de los datos es quien da la aprobación de los modelos generados en la fase anterior.

- Despliegue

  Una vez que ya se cuenta con un modelo y se ha madurado lo suficiente, es necesario dar una presentación adecuada en dónde la representación del conocimiento sea a través de una técnica conocida.

  Una vez que se han descrito de forma general las seis fases de la metodología CRISP - DM para Minería de Datos, es necesario puntualizar en cada una de estas fases cuales son las tareas generales que las integran. Asimismo, es necesario mencionar que la realización de las tareas generales carece de obligatoriedad en algunos casos, ya que CRISP – DM es una metodología abierta, en dónde es permitida la adecuación de tareas con el objeto de ajustar la metodología a la problemática particular.

3.2. Comprensión del negocio

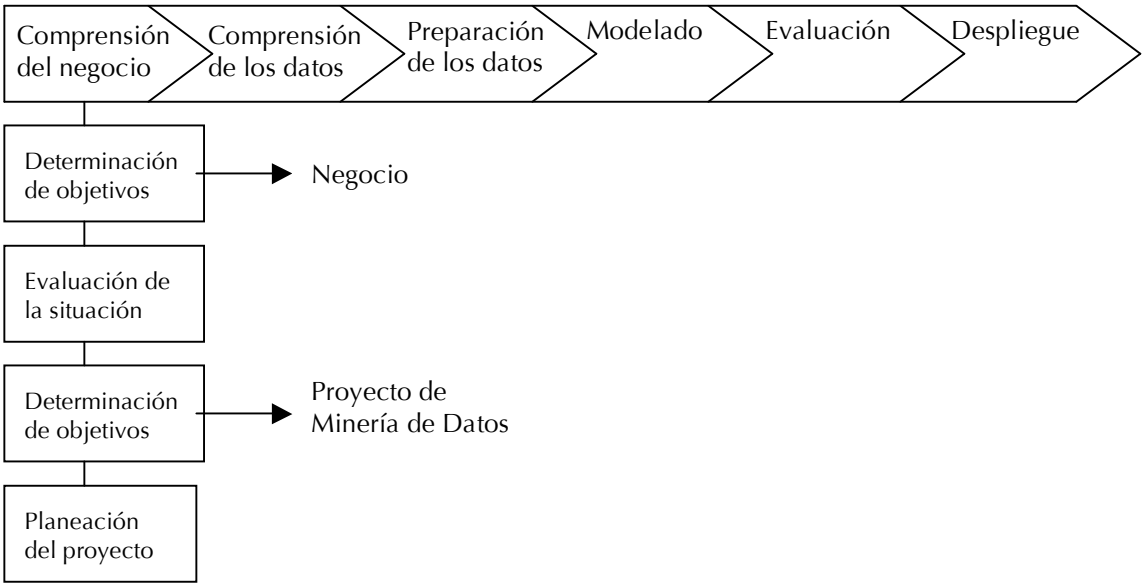

Figura 17. Fase inicial: Comprensión del negocio.

3.2.1. Determinación de los objetivos del negocio

Tarea **Determinación de los objetivos en términos del negocio**

El primer objetivo respecto del análisis de los datos, es su comprensión desde la perspectiva del negocio, es decir hacer una análisis y orientar al cliente sobre lo que realmente desea conseguir con el estudio de Minería de Datos, comprendiendo los objetivos organizacionales y reglas de negocio con las cuales se va a trabajar durante dicho proceso.

La importancia de esta tarea es conocer las inquietudes del cliente, saber qué es lo que busca a través de un estudio de Minería de Datos y realizar el enfoque del proyecto en términos del Modelo de Negocio. Una posible consecuencia de saltar esta tarea es gastar grandes cantidades de tiempo y esfuerzo tratando de producir respuestas correctas a preguntas erróneas.

Salidas[2] **Documento de Contexto**

Es la descripción del modelo de negocio bajo el cual se desenvuelve la organización al principio del proyecto de Minería de Datos.

Objetivos organizacionales

Este documento describe el objetivo principal del proyecto visto desde el punto de vista del cliente, es decir, se plantea la pregunta a la cuál se quiere dar respuesta a través del estudio de Minería de Datos.

[2] Las Salidas son documentos entregables o de referencia para el grupo de trabajo a cargo del proyecto de Minería de Datos.

Perspectiva organizacional sobre los resultados esperados

En este documento se define cual será el entregable final del proyecto, documentación, aplicaciones, formas en que será representado y entregado el conocimiento adquirido durante el proceso.

3.2.2. Evaluación de la situación

Tarea **Evaluación de la situación**

Esta tarea involucra hacer un análisis detallado de la situación actual del negocio con el objeto de determinar los recursos, las restricciones, suposiciones y otros factores que deben ser considerados en la determinación del plan del proyecto. La tarea anterior únicamente debe hacer un análisis preeliminar, para que en esta tarea se conozcan los detalles de la situación actual.

Salidas **Inventario de recursos**

Es la lista de recursos con que se contará durante el proyecto, esto incluye: personal (expertos del negocio, expertos en Minería de Datos, expertos en datos, personal de soporte técnico), datos (información que se va a utilizar, acceso a datos operacionales, históricos, etc.), recursos computacionales (hardware y software y otras herramientas informáticas).

Requerimientos, suposiciones y restricciones

Lista de todos los requerimientos del proyecto, incluyendo la fecha de término del mismo, a su vez, se debe incluir en términos legales las condiciones bajo las cuales se van a operar los datos e información proporcionada por la organización, es decir, las condiciones de confidencialidad de la información.

Lista de todas las suposiciones que se tienen sobre los datos con el objeto de evaluarlas durante el proceso de Minería de Datos.

Lista de las restricciones del proyecto. Este documento plasma los alcances del estudio de Minería de Datos.

Riesgos y contingencias

Lista de los riesgos o eventos que puedan afectar durante el proyecto, llevando a éste a su retraso o fracaso; la lista de riesgos debe incluir un plan de contingencias.

Terminología

Se debe realizar un glosario de los términos más relevantes para el proyecto, éste debe incluir dos componentes:

(1) Terminología relevante del negocio.

(2) Terminología de Minería de Datos.

Costos y beneficios

Análisis costo – beneficio de la realización del proyecto, en dónde sea comparable el costo de llevar el proceso contra los beneficios potenciales que obtendrá la organización.

3.2.3. Determinación de objetivos en términos de Minería de Datos

Tarea **Determinación de objetivos en términos de Minería de Datos**

Este documento plantea en términos técnicos los objetivos del proyecto; por ejemplo un objetivo en términos del negocio sería: "Incrementar las ventas a través de los clientes existentes.", mientras que un objetivo en términos técnicos sería: "Predecir que tan cercanas son las características entre los clientes que adquirieron algún tipo de producto en dónde las características demográficas son edad, salario, ciudad de origen".

Salidas **Objetivos de Minería de Datos**

Se describen los entregables del proyecto que responderán en términos de los objetivos organizacionales.

Criterio de éxito para Minería de Datos

Se describen en términos técnicos los resultados a obtener y el nivel de confianza de los estudios que arrojen como resultado el éxito del proyecto.

3.2.4. Planeación del proyecto

Tarea **Planeación del proyecto**

Documento que describe las líneas de acción a seguir durante el proyecto, especificando tareas, tiempos, responsables y entregables.

Salidas **Plan del proyecto**

Como se ha mencionado, el plan del proyecto debe sustentar todas las actividades a realizar, los responsables, las tareas, los materiales y demás recursos a utilizar para su correcta administración en el proceso de Minería de Datos.

3.3. Comprensión de los datos

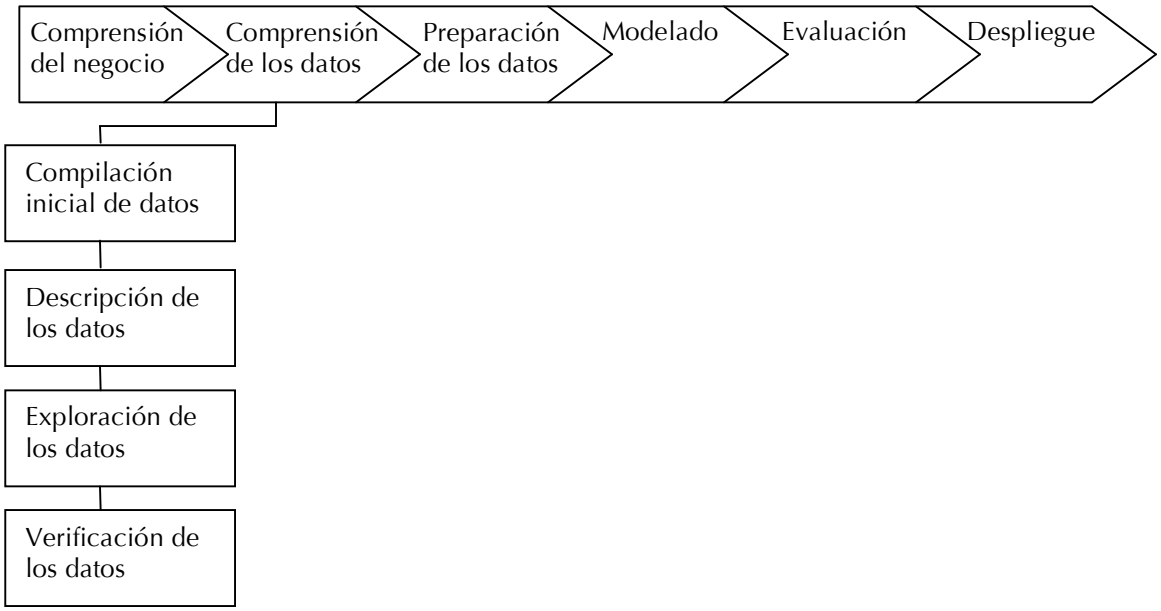

Figura 18. Fase dos: Comprensión de los datos.

3.3.1. Compilación inicial de los datos

Tarea **Compilación inicial de los datos**

Es la obtención de los datos con que se trabajará durante el proyecto, descritos en la lista de recursos, así como el acceso a los mismos en caso de requerir fuentes de datos externas.

Nota: en caso de obtener datos de múltiples fuentes, la integración de los mismos es una tarea adicional.

Salidas **Reporte inicial sobre la adquisición de los datos**

Elaborar un reporte que contenga cuales son las fuentes de datos a utilizar, la forma y cantidad de los mismos así como reportar cuales son y en el caso de haberlos, los problemas para acceder a la información.

3.3.2. Descripción de los datos

Tarea **Descripción de los datos**

El propósito de esta tarea es el examinar los datos con el objeto de determinar la superficialidad ó grosor de los datos.

Salidas **Reporte sobre la descripción de los datos**

La descripción de los datos incluye: el formato, la cantidad y calidad de los mismos, con el objeto de determinar si son útiles para el estudio.

3.3.3. Exploración de los datos

Tarea **Exploración de los datos**

Es el primer análisis que se hace de los datos, se utilizan técnicas para conocer los datos, la relación entre las variables a analizar, también se determina el atributo o atributos que serán clave en el estudio.

Salidas **Reporte de la exploración de los datos**

Describe los resultados obtenidos en la descripción de los datos incluyendo la hipótesis inicial y su impacto en la estructura actual del proyecto. Es recurrente incluir gráficos y técnicas estadísticas descriptivas que permitan el conocimiento de los datos a manejar.

3.3.4. Verificación de los datos

Tarea **Verificación de los datos**

En esta tarea se examina la calidad de los datos, en donde se responden preguntas referentes a si los datos cumplen con las características requeridas, si satisfacen todos los casos necesarios, si existen valores nulos, o si existe alguna anomalía.

Salidas **Reporte de calidad de los datos**

Lista de los resultados de la tarea de verificación de datos; si se determina la existencia de problemas con los datos, entonces, se deberá listar las posibles soluciones.

3.4. Preparación de los datos

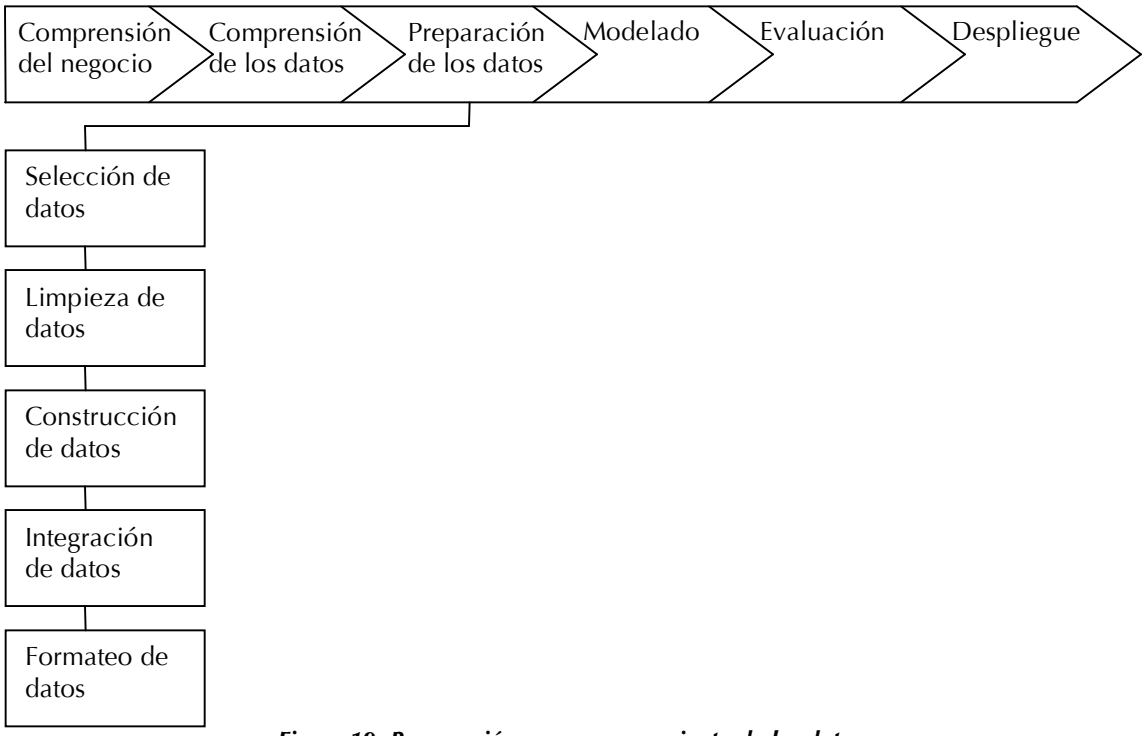

Figura 19. Preparación o preprocesamiento de los datos.

Para dar un sentido más amplio dentro de la explicación de tareas referente a esta fase de la metodología, es necesario, en primer término, conocer cuáles serán las salidas al final de la fase, es decir los objetivos que se deberán alcanzar.

Salidas **Conjunto de datos a utilizar en el proyecto**

Como resultado de la preparación de datos se debe obtener un conjunto de datos lo suficientemente sólido para que sea utilizado durante todo el proyecto, ya que éste servirá para generar el modelo de Minería de Datos que dé explicación al fenómeno de estudio.

Descripción del conjunto de datos a utilizar en el proyecto

Describe el conjunto de datos que serán utilizados en el proyecto.

3.4.1 Selección de datos

Tarea **Selección de datos**

En esta tarea se determina cuales serán los datos que serán incluidos para su análisis. El criterio a utilizar para discriminar los datos deberá estar basado en la relevancia de los objetivos de Minería de Datos y en las restricciones impuestas por los propietarios de los datos, así como el volumen de datos con que se cuente y el número de variables existentes.

Salidas ***Determinación para la inclusión/exclusión de datos***

Lista de datos que serán incluidos y excluidos del proyecto de Minería de Datos, así como las razones por la cuales se incluyeron o excluyeron los mismos.

3.4.2. Limpieza de datos

Tarea **Limpieza de datos**

La limpieza de datos deberá ser entendida como el proceso de incrementar la calidad de los datos para alcanzar el nivel requerido por las técnicas de análisis previamente seleccionadas. Esta tarea involucra la homogenización de valores nulos, inserción de valores por defecto y estimación de datos.

Salidas **Reporte sobre la limpieza de datos**

Este documento describe cuáles fueron las decisiones y acciones tomadas durante la tarea de limpieza de datos; es importante mencionar que el objetivo principal de esta tarea se cumple cuando se satisfacen las necesidades expuestas durante la tarea de ***verificación*** en la fase de ***Comprensión de los datos***.

3.4.3. Construcción de datos

Tarea **Construcción de datos**

Esta tarea incluye operaciones para la construcción de atributos derivados, generación de nuevos registros o transformación de valores existentes que permitan enriquecer el conjunto de datos original con el objeto de satisfacer las necesidades requeridas por el modelo de Minería de Datos a construir.

Salidas **Atributos derivados**

Son aquellos atributos que se generan a través de la realización de operaciones entre atributos existentes en el conjunto de datos original. Ejemplo: Monto Parcial = Cantidad * Precio Unitario.

Registros generados

Describe la generación de nuevos registros en el conjunto de datos original. Ejemplo: Inclusión de nuevos registros de ventas para los clientes que no realizaron compras durante el año pasado. Lógicamente no tendría sentido incluir registros que indiquen que los clientes no realizaron compras, sin embargo, esto se realiza con el propósito de ser suficientemente explícitos al momento de indicar que el cliente no realizó compras el año pasado.

3.4.4. Integración de datos

Tarea **Integración de datos**

Consiste en la aplicación de un conjunto de métodos para combinar información proveniente de diversas fuentes.

Salidas **Conjunto de datos integrados**

Debido a que la información puede provenir de diversas fuentes, ésta deberá ser integrada en un mismo conjunto de datos el cuál será la salida de información. La principal forma de integrar datos en una base de datos relacional es a través del operador `JOIN`, el cuál permite conjuntar información de diversas tablas.

3.4.5. Formateo de datos

Tarea **Formateo de datos**

La tarea de dar formato a los datos consiste primordialmente en realizar modificaciones a los datos que permitan satisfacer el orden en que éstos deberán ser procesador por la herramienta de Minería de Datos seleccionada.

Salidas **Conjunto de datos con formato**

Es importante conocer la herramienta de Minería de Datos a utilizar durante el proyecto, ya que, durante la construcción del modelo es posible que la entrada de datos tenga un formato u orden específico para obtener los resultados adecuados.

Dicho formato de datos puede consistir en ordenar los atributos de acuerdo a un atributo en particular, delimitar las cadenas de texto del conjunto de datos para ajustarlas a un tamaño en particular, etc.

3.5. Modelado

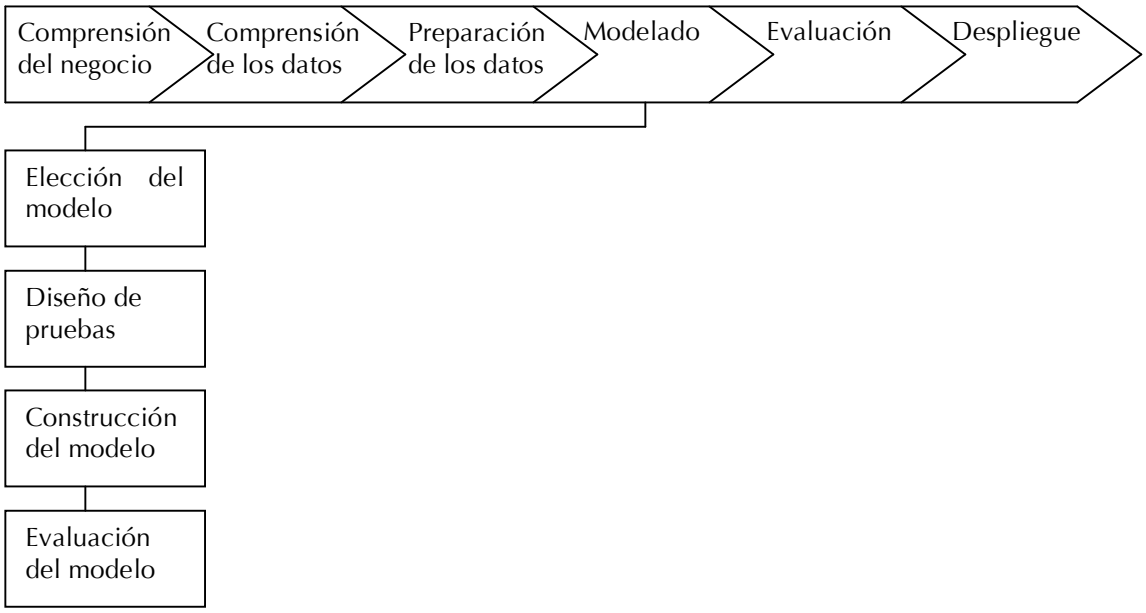

Figura 20. Fase de Modelado de datos.

3.5.1. Elección del modelo

Tarea **Elección de la técnica de modelado**

Como primera tarea de la fase de modelado está la elección de la técnica de modelado a utilizar, es decir la elección de la herramienta que servirá para la construcción de un modelo de Minería de Datos que dé explicación al fenómeno del modelo de negocio; esta tarea debe realizarse debido a que un mismo fenómeno puede ser explicado a través de diversos caminos, por ejemplo, si se requiere generar un árbol de decisiones, éste puede construirse a través del algoritmo C4.5 o bien con la generación de una red neuronal de propagación.

Salidas **Definición del modelo a utilizar en el proyecto**

Documentar la técnica o mejor dicho la herramienta de modelado a utilizar en el proyecto de Minería de Datos.

3.5.2. Diseño de pruebas

Tarea **Diseño de pruebas**

Antes de comenzar con la construcción de la herramienta de modelado es necesario generar un procedimiento o mecanismo que compruebe la calidad y viabilidad de la misma.

Por ejemplo, en la construcción de herramientas para la clasificación de datos es común la creación de registros erróneos que servirán de medidas de calidad, a su vez, típicamente el conjunto de datos a utilizar es dividido en dos partes una para entrenamiento del modelo y otra para comprobación del modelo.

Salidas **Diseño de pruebas**

Generación de un documento que describa el procedimiento a utilizar para la prueba y evaluación del modelo de Minería de Datos a generar a través de una herramienta en particular. El componente primario del plan de pruebas consiste en decidir como es que se dividirá el conjunto de datos a utilizar.

3.5.3. Construcción del modelo

Tarea **Construcción del modelo**

Una vez que se ha determinado la herramienta de Minería de Datos a utilizar ésta deberá ser construida y aplicada sobre el conjunto de datos preparado con el objeto de generar uno o más modelos explicativos del modelo de negocio.

Salidas **Configuración de parámetros**

Sea cual sea la herramienta de modelado que se utilice será necesario configurar el conjunto de parámetros necesarios para trabajar.

Modelos

Es el conjunto actual de modelos generado por la herramienta utilizada.

Descripción del modelo

Documento que describe el modelo resultante de la aplicación de la herramienta Minería de Datos en el conjunto de datos del proyecto. El reporte incluye la interpretación del modelo y la documentación de cualquier dificultad presentada durante la utilización de la herramienta construida.

3.5.4. Evaluación del modelo

Tarea **Evaluación de modelo**

Una vez que se ha generado un modelo, éste deberá ser interpretado por los expertos de Minería de Datos, quienes utilizarán su criterio y el conjunto de pruebas diseñado para evaluar si el resultado es confiable, en dicho caso se deberá contactar a los analistas y dueños del negocio para discutir los resultados obtenidos bajo un contexto menos técnico y más del modelo de negocio.

Salidas **Evaluación del modelo**

Documento que enlista los resultados obtenidos en esta tarea, dichos resultados deberán estar organizados de acuerdo al nivel de importancia que reflejen.

Revisión de los parámetros

En caso de que el modelo generado no sea lo suficientemente maduro, se deberán ajustar los parámetros ingresados a la herramienta construida con el fin de iterar las veces necesarias hasta obtener resultados satisfactorios.

3.6. Evaluación

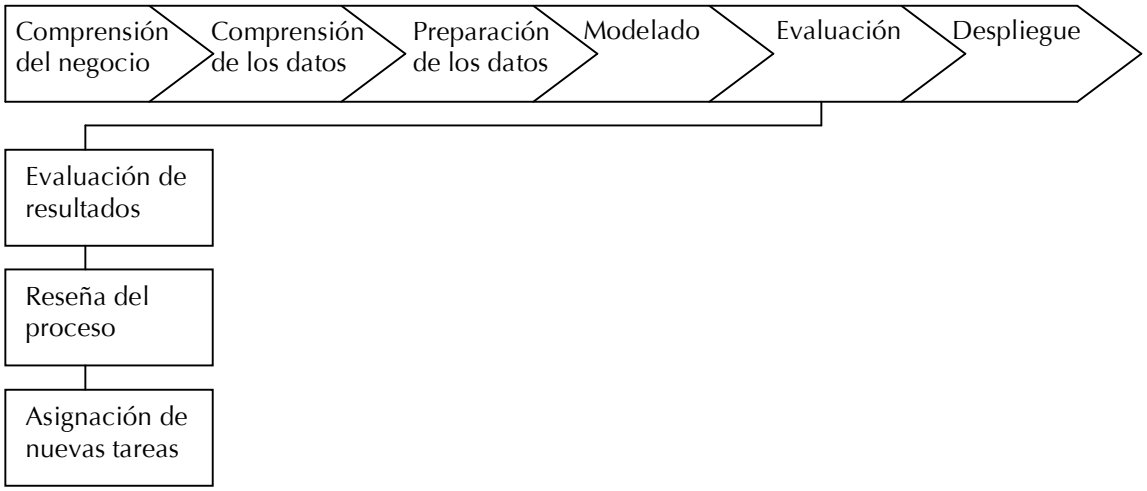

Figura 21. Fase de evaluación.

3.6.1. Evaluación de resultados

Tarea **Evaluación de resultados**

Esta tarea consiste en la evaluación de resultados por parte de los expertos del negocio quienes determinan la aceptación de resultados, o en su caso, las deficiencias encontradas en los resultados presentados.

Otra opción para evaluar los resultados consiste en la aplicación del conocimiento generado.

Los resultados obtenidos deberán satisfacer completamente el objetivo planteado por los expertos del negocio en la etapa de ***determinación de objetivos***.

Salidas **Evaluación de los resultados de Minería de Datos respecto del criterio de los expertos del negocio**

Documento que describe los criterios que fueron utilizados para evaluar los resultados obtenidos, incluyendo una comparativa entre los resultados obtenidos y los objetivos planteados al principio del proyecto.

Modelos aprobados

Luego de que se evaluaron los modelos obtenidos como resultado del proceso de negocio y que estos han sido aprobados de acuerdo al criterio de los expertos del negocio, los modelos se convierten en modelos aprobados, los cuales deberán ser documentados.

3.6.2. Reseña del proceso

Tarea **Reseña del proceso**

Una vez que se han evaluado los modelos, será necesario dar una reseña a los expertos del negocio sobre el proceso seguido, detallando los datos utilizados y la interpretación que se hizo de los mismos con el objeto de que se señale, si es que fuese necesario, si se han omitido características de los datos o se han realizado interpretaciones erróneas.

Salidas **Reseña del proceso**

Documento que resume en términos claros las actividades que deberán ser mejoradas o repetidas con el objeto de generar un nuevo modelo que contemple las omisiones realizadas en los datos.

3.6.3. Asignación de nuevas tareas

Tarea **Asignación de nuevas tareas**

Acorde a los resultados obtenidos se deberá determinar el estado del proyecto de Minería de Datos con la finalidad de establecer las tareas subsecuentes como podría ser la finalización del proyecto, el despliegue del mismo o bien la determinación de nuevos objetivos.

Salidas **Lista de posibles acciones**

Lista de las acciones que se deberán seguir a fin de concluir el proyecto y/o comenzar con uno nuevo.

Decisión

Describir la forma de proceder ante las nuevas tareas asignadas o ante la finalización del proyecto.

3.7. Despliegue

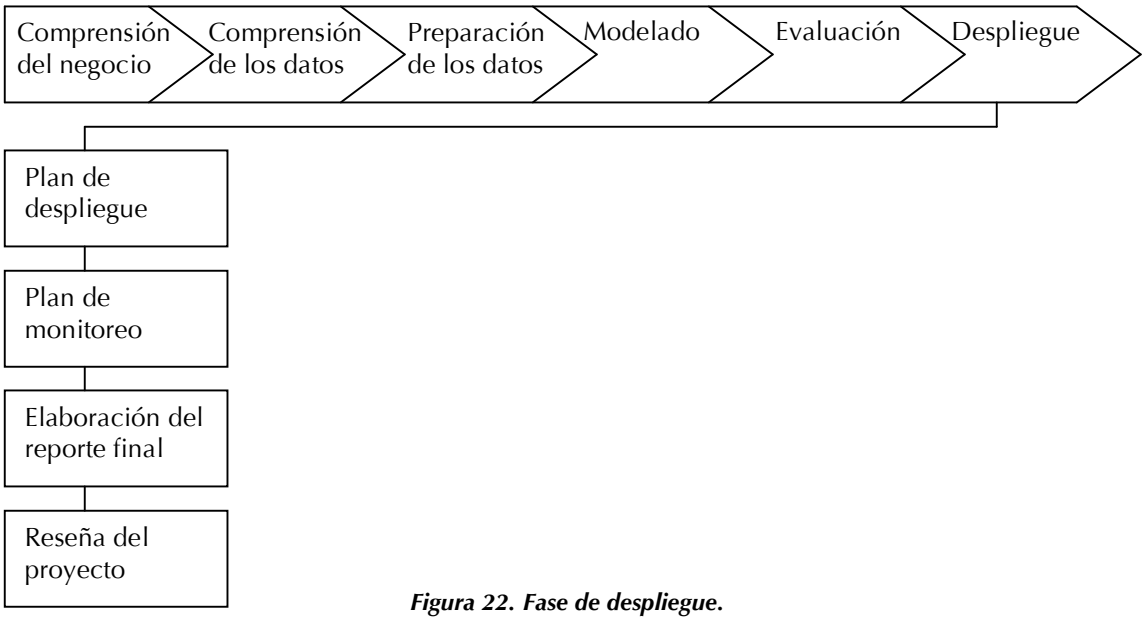

Figura 22. Fase de despliegue.

3.7.1. Plan de despliegue

Tarea **Plan de despliegue**

En esta tarea se determina el plan a seguir para la aplicación del conocimiento generado durante el proceso de Minería de Datos, entendiendo que primeramente el modelo fue evaluado y aprobado tanto por los expertos de Minería de Datos como por los expertos del negocio.

Salidas **Plan de despliegue**

Documento que resume las estrategias a seguir para el despliegue del modelo generado.

3.7.2. Plan de monitoreo y mantenimiento

Tarea **Plan de monitoreo y mantenimiento**

Esta tarea involucra la creación de un plan de monitoreo y mantenimiento que en primer término evalué constantemente los resultados obtenidos en la aplicación del conocimiento generado, a su vez se deberá contemplar que un sistema siempre es cambiante y por ello en su momento se deberá dar mantenimiento a la herramienta de Minería de Datos generada para que los resultados arrojados siempre sean los idóneos.

Salidas **Plan de monitoreo y mantenimiento**

Documento que describe las estrategias a seguir para el monitoreo y mantenimiento de la herramienta de Minería de Datos generada.

3.7.3. Elaboración del reporte final

Tarea

Al final del proyecto, el administrador del mismo deberá elaborar un reporte final el cuál deberá incluir una reseña sobre las experiencias adquiridas durante el proyecto, así como una descripción en términos comprensivos sobre la presentación de resultados.

Salidas **Reporte final**

Documento que describe sobre las experiencias adquiridas durante el proyecto, así como una descripción en términos comprensivos sobre la presentación de resultados. También se deberán anexar los documentos generados a lo largo del proyecto.

Presentación final

Se deberá realizar una presentación formal ante los dueños del negocio con el objeto de presentar los resultados del proyecto.

3.7.4. Reseña del proyecto

Tarea **Reseña del proyecto**

La reseña del proyecto incluye una descripción sobre las dificultades presentadas, las tareas realizadas y las actividades que fueron mejoradas a lo largo del proyecto.

Salidas **Reseña del proyecto**

Documento que describe las experiencias adquiridas durante el proyecto.

Capítulo 4

Enfoque práctico de la Minería de Datos

Una vez que se tiene como trasfondo cuáles son las necesidades históricas de almacenar información y se ha determinado que en la actualidad cuando se presentan grandes volúmenes de datos es posible realizar análisis exhaustivos sobre ellos con el objeto de generar conocimiento que sea utilizado para la evolución de las organizaciones, es como se plantea un caso de estudio, en dónde se justificará, a través de la práctica, que es posible generar conocimiento, tomando como base grandes volúmenes de información y a través de un proceso de Minería de Datos.

Este capítulo, por tanto, comenzará con un apartado sobre el papel que juega la Minería de Datos en los negocios, así como la importancia de los sistemas de información en el proceso de Minería de Datos; posteriormente se llevará acabo un proceso de Minería de Datos como caso práctico, el cuál estará guiado por la metodología CRISP – DM 1.0, misma que fue descrita en el capítulo anterior.

Así pues, con esta breve introducción es como damos paso al contenido del capítulo.

4.1. La Minería de Datos en los negocios

Si bien la Minería de Datos puede darse dentro de cualquier ámbito en dónde se cuente con grandes volúmenes de información, se ha decidido para efecto práctico restringir el campo de aplicación en éste capítulo, en donde nos avocaremos a las organizaciones empresariales, es decir, organizaciones que persiguen fines de lucro y en dónde se ofrecen productos y/o servicios a clientes.

En el ámbito empresarial, las principales aplicaciones que tiene la Minería de Datos se ven reflejadas en la optimización de los procesos de marketing y ventas, así como en administración de la relación con el cliente (Customer Relationship Management, CRM por sus siglas en inglés).

Para la Minería de Datos en los negocios, siempre surgen retos como: generación de perfiles de clientes basados en conductas de consumo, generación de grupos de productos que ofrecen mayores ganancias, generación de árboles de decisión en dónde se determinen sociedades de productos adquiridos, generación de pronósticos de ventas, creación de políticas para el control de inventarios, etc.

La mayoría de las empresas, hoy en día, avocan la mayor parte de sus esfuerzos y estrategias pensando en los productos que ofertan, en donde un factor crítico de éxito es la innovación de productos provistos con calidad y funcionalidad, sin embargo frecuentemente olvidan el servicio previo que hay que ofrecer a los clientes para que estos decidan realizar una inversión en sus productos, ya que no importa lo efectivo que sea un producto si no se establece el ambiente adecuado para que éste sea adquirido por una persona.

> *"La principal idea de la Minería de Datos enfocada al CRM es hacer útil la información histórica de las compras de los clientes, con el objeto de utilizarla estratégicamente en el futuro. Esto es posible siempre y cuando sea almacenada la conducta de consumo de las personas, ya que ésta reflejará diferentes necesidades, preferencias, aficiones e intereses de los clientes." (Berry y Linoff 2004: 6)*

En los negocios la Minería de Datos tiene como objetivo primordial realizar análisis que sean de utilidad al momento de la toma de decisiones que contribuyan a una mayor captación de ingresos monetarios.

> *"La Minería de Datos es una herramienta y como cualquier otra herramienta, no es suficiente comprender su funcionamiento, sino que, es necesario comprender como ésta puede ser utilizada." (Berry y Linoff 2004: 6)*

4.2. El rol de los sistemas de información operacionales

Las organizaciones sean del tipo de que sean, hoy en día implementan sistemas de información con el objeto de dar soporte a las actividades que realizan, ya que éstos les permiten almacenar el detalle de sus operaciones para que en un futuro sirvan de base para la generación de reportes gerenciales que resuman la información de manera tal que un fenómeno pueda ser observado de manera más amplia y no sobre el detalle de la situación.

El principal rol de los sistemas de información operacionales en una organización es la captación de los datos, y por ende, la captación de los mismos debe realizarse de la forma más simple, eficiente y sin dar pie a la generación de información basura o a la omisión de captura de datos.

Los sistemas de información operacionales pueden ser vistos como la principal fuente de información que será utilizada en un proceso de Minería de Datos.

4.3. El rol del Data Warehouse

> *"Data warehouse: Conglomeración de los de datos de una organización y áreas de presentación en un almacén, en dónde los datos operacionales son estructurados específicamente para la realización de consultas y análisis, de modo en que los datos posean un gran rendimiento y sean fáciles de utilizar y consultar." (Kimball y Ross 2002: 397)*

> *"Data warehouse: es una colección de datos orientadas a un dominio, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones de una empresa u organización. Se trata, sobre todo, de un expediente de una empresa, más allá de la información transaccional y operacional, almacenando la información en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente OLAP, procesamiento analítico en línea)." (Web. 03)*

Las compañías que enfocan sus esfuerzos en la captación de recursos a través del conocimiento de sus clientes, realizan una infinidad de estudios con el objeto de aprender de lo que el cliente desea para después ofrecer los productos adecuados a sus necesidades; la manera en que una compañía adquiere información relacionada a los clientes, es a través de encuestas, sistemas de información operacionales, registros de llamadas a los centros de atención a clientes, y otros estudios de mercadotecnia; sin embargo, es difícil adquirir conocimiento de los clientes teniendo diversas fuentes de datos en donde a su vez, éstos se encuentran sin orden que permita la realización de estudios especializados.

> *"Muchas compañías a lo largo de su historia, generan y adquieren terabytes o gigabytes de información de y acerca de sus clientes, productos y entorno, sin que aprendan nada de ella. Los datos son generados en principio debido a que se busca un propósito operacional como el control de inventarios, administración del proceso de ventas para conocer el estado de pedidos, etc.*
>
> *Para que la generación de conocimiento tenga lugar, es necesario tomar como base la información proveniente de diversas fuentes, conjuntarla y organizarla de forma consistente y útil, haciendo una limpieza de la misma para que ésta pueda ser entendible y útil en un proceso de generación de conocimiento. A esto se le conoce como data warehousing." (Berry y Linoff 2004: 4-5)*
>
> *"Un buen data warehouse provee el acceso a la información generada por un sistema de información operacional en donde el formato de los datos sea mucho más amigable que en la forma en que los datos son almacenados por el sistema de información que los genera. Idealmente los datos existentes en un data warehouse deben haber sido preprocesados, es decir que los datos han sido integrados de diversas fuentes, limpiados, unidos, igualados respecto a un conjunto de registros similares en caso de contar con valores faltantes y sumarizados en diversas formas útiles." (Berry y Linoff 2004: 4-6)*

El rol que juega el data warehouse dentro de un proceso de Minería de Datos, consiste básicamente en proveer el conjunto de datos a utilizar una vez que estos han sido integrados, preprocesados y organizados de una forma útil para el estudio a realizar.

Capítulo 5

Minería de Datos: Aplicación en un caso práctico

El caso práctico a desarrollar en el presente documento de tesis se enfoca a la realización de un estudio de Minería de Datos para la Dirección de Computo para la Docencia (DCD) de la Dirección General de Servicios de Cómputo Académico (DGSCA) de la Universidad Nacional Autónoma de México (UNAM), en donde el trabajo realizado permita confirmar o negar la hipótesis de que la Minería de Datos puede intervenir como una herramienta para la toma de decisiones en el proceso de calendarización de cursos de cómputo.

El desarrollo del caso práctico se encuentra guiado por la metodología CRISP – DM 1.0, explicada en capítulos anteriores, así pues comenzaremos este capítulo con la comprensión del modelo de negocio que sigue la DCD, de la DGSCA, UNAM.

5.1. Comprensión del negocio

La Dirección General de Servicios de Cómputo Académico de la UNAM es la entidad universitaria encargada de la operación de los sistemas centrales de cómputo académico y de las telecomunicaciones de la institución; su esfuerzo más amplio es la capacitación en tecnología de la información, de prospección e innovación y de asimilación de estas tecnologías en beneficio de la Universidad y de la sociedad en general.[3]

En contexto con los objetivos anteriores, la DGSCA cuenta con la Dirección de Cómputo para la Docencia (DCD) la cuál está encargada de la capacitación y actualización en tecnología de la información al público en general, a través de una serie de cursos, líneas de especialización y diplomados.

La DGSCA cuenta con un conjunto de Centros de Extensión, los cuales se encargan de dar soporte a las actividades que persigue la DCD. Los centros de extensión se encuentran descentralizados debido a la extensión geográfica que ocupan, pero el funcionamiento interno es siempre el mismo en cada uno de ellos, teniendo la información centralizada en la sede principal: DGSCA – CU. La estructura orgánica en cada uno de los centros de extensión está conformada entre otros por:

- Departamento de Informes y Relaciones
- Departamento de Control Escolar
- Departamento de Infraestructura

[3] **Pisanty Baruch, Alejandro**. *Quienes somos*, en http://www.dgsca.unam.mx/somos.html. Visitada el 15 de febrero de 2007.

La Dirección de Cómputo para la Docencia, a través de la Subdirección de Planeación Académica, administra y mantiene el Sistema de Administración de Educación Continua (SAEC), que como su nombre lo indica, es un sistema diseñado para la administración de información de control escolar que generan los Centros de Extensión de la Dirección de Cómputo para la Docencia de la DGSCA. En este sentido, en el SAEC se registran los datos de todos y cada uno de los cursos, líneas de especialización y diplomados que son impartidos en los Centros de Extensión, así como la información de alumnos que se inscriben a dichos cursos, los profesores que los imparten, aulas donde son impartidos, etc.

Como parte de las actividades de la Subdirección de Planeación Académica, está el realizar una adecuada calendarización de los cursos, líneas de especialización y diplomados, que permita una adecuada captación de público a través de la adecuada distribución de cursos en las Sedes de la DGSCA con el objeto de allegarse de un mayor número de recursos y disminuir el número de cursos programados sin impartir por parte de la Institución.

A lo largo de los años, el SAEC ha constituido una fuente de información con la capacidad de ayudar al soporte operacional de la Dirección de Cómputo para la Docencia de la DGSCA, la cantidad de información colectada y presentada por este sistema de información tiene como repercusión grandes volúmenes de datos que son apreciados de forma particular, haciéndose difícil de apreciar en algunos casos, de forma conjunta o en grupos más sólidos el fenómeno bajo el cual se desenvuelve el Modelo de Negocio de la Institución; siendo esto, más que una desventaja, una oportunidad ya que el volumen de datos generados por el SAEC cuenta con la potencialidad necesaria para la aplicación de técnicas de Minería de Datos que permitan generar conocimiento que a su vez sea utilizado como una ventaja competitiva al momento de realizar la calendarización de cursos, líneas de especialización y diplomados que imparte la DGSCA.

5.1.1. Determinación de los objetivos del negocio

5.1.1.1. Objetivos organizacionales

Identificar de forma clara cuál es el posicionamiento de los cursos que son impartidos en los centros de extensión de la DGSCA, con el objeto principal de realizar una optima calendarización de cursos de cómputo.

Allegarse de un mayor número de recursos y disminuir el número de cursos programados sin impartir, a través de una adecuada calendarización de cursos por parte de la Institución.

5.1.1.2. Perspectiva organizacional sobre los resultados esperados

Al finalizar el proyecto de Minería de Datos, se deberá presentar el proyecto a las autoridades de la dependencia, haciendo entrega de los materiales informáticos generados durante el estudio (Programas, librerías, consultas, etc.). Asimismo, se deberá presentar la documentación técnica del proyecto, explicando de forma detallada, cada uno de los gráficos y tablas contenidas en el mismo.

5.1.2. Evaluación de la situación

5.1.2.1. Inventario de recursos

<table>
<tr><th>Humanos</th><th>Datos a Utilizar</th></tr>
<tr><td>Expertos del Negocio</td><td rowspan="3">Información histórica sobre los cursos que son impartidos, obteniendo a detalle la situación presentada en las inscripciones a los cursos.</td></tr>
<tr><td>LI. Juana Figueroa Reséndiz
Coordinadora de Producción Académica</td></tr>
<tr><td>Mtro. Jesús Díaz Barriga Árceo
SubDirector de Planeación Académica</td></tr>
<tr><td>Expertos de Minería de Datos</td><td>Acceso a Datos</td></tr>
<tr><td>Carlos Tomás Reyes García
Pasante de la Lic. En Informática</td><td>Se requerirá el acceso a la base de datos que utiliza el sistema SAEC, con el objeto de recuperar la información de los cursos impartidos.</td></tr>
<tr><td colspan="2">Recursos computacionales</td></tr>
<tr><td>Equipo de cómputo</td><td>Software</td></tr>
<tr><td>▪ Windows XP (Service Pack 1 or 2)
▪ Intel Pentium IV o superior
▪ Disco Duro 40 GB
▪ Memoria RAM 512 MB</td><td>▪ Suite de Microsoft Office 2003 (Word, Excel, Visio, PowerPoint, Project)
▪ SQL Server 2000
▪ MatLab R13
▪ Weka 3.4
▪ Cristal Reports XI
▪ Adobe Acrobat Profesional</td></tr>
</table>

5.1.2.2. Requerimientos, suposiciones y restricciones

Requerimiento

Generación de un indicador de cursos que permita ver la efectividad de un curso considerando diversos factores.

Restricciones

El listado de cursos que arrojará el indicador antes construido, deberá presentarse de acuerdo a la información particular de cada una de las sedes.

Suposiciones

Todas las sedes presentarán resultados similares.

5.1.2.3. Convenio de confidencialidad de la información a utilizar

Ciudad Universitaria, a 2 de julio de 2007.

Mat. Carmen Bravo Chaveste.
Directora de Cómputo para la Docencia
DGSCA, UNAM

P r e s e n t e

Por medio del presente documento permítame hacer de su conocimiento que me encuentro laborando en la Coordinación de Producción Académica de la Subdirección de Planeación, en donde como actividades principales tengo encomendada la administración y mantenimiento operacional del Sistema de Administración de Educación Continua (SAEC).

Con el objeto de realizar una mejora en el proceso de calendarización de cursos de cómputo que imparte esta dependencia, presenté meses atrás una propuesta, la cuál consiste en la realización de un estudio de Minería de Datos sobre la base de datos del sistema SAEC, en donde el resultado obtenido por dicho estudio sirva como herramienta para la toma de decisiones en el proceso antes señalado.

Es importante mencionar que el estudio de Minería de Datos propuesto a esta dependencia, forma parte del documento de tesis que me encuentro desarrollando como opción de titulación a la Licenciatura en Informática de FCA de la UNAM. Dicho documento de tesis lleva el titulo "La Minería de Datos como Herramienta para la Toma de Decisiones en el Proceso de Calendarización de Cursos de Cómputo", la cual incluye un capítulo donde se presenta un caso de aplicación práctico, que estará basado en el estudio propuesto a la Subdirección de Planeación.

Por lo anterior solicito su aprobación para publicar, en el documento de tesis que me encuentro desarrollando, un conjunto de registros históricos de la base de datos de SAEC, en donde se muestra información referente a cursos impartidos. El caso de aplicación práctico tiene como objetivo evaluar la efectividad de los cursos tomando en cuenta 10 factores, que se analizarán multivariadamente con el objeto de reducir la dimensión del problema de 10 a 1 variable que explique la efectividad de los cursos impartidos por la DGSCA.

Los datos extraídos comprenden un resumen histórico con los siguientes datos:

Nombre del curso	Los campos de resumen que se utilizarán para el estudio, abarcarán información contenida en la base de datos de SAEC para un periodo comprendido entre 1 de enero de 2006 al 15 de junio de 2007 La propuesta de confidencialidad consiste en ocultar los datos referentes a cantidades monetarias, ocultando la información real al lector.
Número de veces impartido	
Alumnos Atendidos	
Alumnos Aprobados	
Alumnos Reprobados	
Cuota	
Honorarios de Profesores	
Gastos Administrativos	
Recuperación Mínima	
Ingresos	
Costo de las Becas	
Monto Neto	
Utilidad	

Cabe resaltar que el conjunto de datos a utilizar no se publicará como tal, sino que, se utilizarán los datos para generar reportes estadísticos y gráficos que resuman la información. La totalidad de los datos servirá para la realización de operaciones estadísticas y matemáticas que permitan obtener una visión general de la información, más no su detalle.

Dada la importancia de la información que utilizaré durante dicho estudio me comprometo a tratarla de manera confidencial, sin que afecte los fines que persigue ésta institución.

Así pues, solicito su autorización para mostrar los datos que usted crea conveniente en el documento de tesis antes mencionado.

Sin más por el momento me despido, quedando de usted en la mejor disposición.

Atte. Carlos Tomás Reyes García.
Coordinación de Producción Académica
DCD, DGSCA, UNAM

5.1.2.4. Aprobación sobre la publicación del proyecto

Universidad Nacional Autónoma de México
Secretaría General
Dirección General de Servicios de Cómputo Académico

DIRECCIÓN DE CÓMPUTO PARA LA DOCENCIA
SUBDIRECCIÓN DE PLANEACIÓN ACADÉMICA
Oficio Núm.: DGSCA/DCPD/SPA/126 /2007

Asunto: Aprobación contenido de tesis

C. Carlos Tomás Reyes García
Presente

Por medio de este conducto le informamos que Luego de haber revisado el contenido del documento de tesis que desarrolló bajo el título: "La Minería de Datos como Herramienta para la Toma de Decisiones en el Proceso de Calendarización de Cursos de Cómputo", se ha determinado la aprobación por parte de esta Subdirección para su publicación bajo los términos de confidencialidad establecidos de común acuerdo, en el cual usted se compromete a no revelar ni publicar información confidencial de esta dependencia.

Sin otro particular, aprovecho la ocasión para enviarle un cordial saludo.

Atentamente
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, D. F., a 13 de agosto de 2007.

M. en C. Jesús Díaz Barriga Arceo
Subdirector de Planeación Académica

Copia: Mat. Carmen Bravo Chaveste, Directora de Cómputo para la Docencia, DGSCA, UNAM

JDB/MCH

Circuito Exterior, Ciudad Universitaria
www.dgsca.unam.mx

5.1.3. Determinación de objetivos en términos de Minería de Datos

5.1.3.1. Objetivo en términos de Minería de Datos

Generar un análisis de componentes principales sobre datos existentes en el fenómeno de calendarización e impartición de cursos de cómputo, en donde dado un conjunto de variables que caracterizan una situación se realice un análisis multivariado de datos.

Evaluar la efectividad de los cursos a través de un estudio de Componentes Principales tomando en cuenta 12 factores que se analizarán multivariadamente con el objeto de reducir la dimensión del problema de 12 a 1 variable que explique la efectividad de los cursos impartidos por la DGSCA.

5.1.3.2. Criterio de éxito

El criterio de éxito para el estudio de componentes principales, consiste en obtener un nivel de explicación del problema en términos de varianza donde se alcance un nivel de confianza igual o superior al 90%; es decir, los componentes principales generados durante el estudio de Minería de Datos deberán explicar el fenómeno con un grado de confianza del 90%.

5.1.4. Planeación del proyecto

5.1.4.1. Alcance

Bajo el marco de referencia anterior y durante los próximos meses, se realizará un estudio de Minería de Datos que permita obtener nuevo conocimiento que aporte a la DGSCA las bases necesarias para aplicar técnicas y estrategias para la calendarización de los cursos que son impartidos a través de la Dirección de Cómputo para la Docencia.

La propuesta va encaminada a utilizar técnicas de Minería de Datos bajo un Modelo Descriptivo para realizar un Análisis de Clusters[4] que permita agrupar los diferentes cursos que imparte la DGSCA y determinar su efectividad en términos sencillos.

De manera más específica la técnica de Minería de Datos a utilizar en este proyecto será la técnica de Análisis de Componentes Principales.

5.1.4.2. Objetivo del proyecto

Incrementar los niveles de eficiencia en el proceso de calendarización de cursos de la DGSCA a través de un conjunto de técnicas de Mineria de Datos aplicadas sobre la base de datos del SAEC, con el objeto de allegarse de un mayor número de recursos y disminuir el número de cursos programados sin impartir por parte de la Institución.

---

[4] El Análisis Clusters, también conocido como Análisis de Conglomerados, Taxonomía Numérica o Reconocimiento de Patrones, es una técnica estadística multivariante cuya finalidad es dividir un conjunto de objetos en grupos (cluster en inglés) de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y los de los objetos de clusters diferentes sean distintos (aislamiento externo del grupo).

5.1.4.3. Actividades a realizar

El proyecto se llevará acabo en un lapso de cuatro meses en dónde las actividades a contemplar incluyen las de la metodología CRISP –DM v1.0 para el desarrollo de un Proyecto de Minería de Datos, así como las actividades pertinentes para la administración de un proyecto. De manera general se plantea el conjunto de actividades a realizar:

Actividades	Responsable(s)	Fecha de inicio	Objetivo
1. Planeación del Proyecto	Carlos Reyes García Líder del proyecto	Febrero 1 de 2007	Contemplar las actividades a realizar, determinación de objetivos y metas del proyecto con el fin de determinar una línea efectiva de trabajo.
2. Comprensión del Negocio	Carlos Reyes García Líder del proyecto	Febrero 19 de 2007	Identificar de forma general el Modelo de Negocio bajo el cuál se desenvuelve la Dirección de Cómputo para la Docencia de la DGSCA.
3. Comprensión de los datos	Carlos Reyes García Líder del proyecto	Febrero 19 de 2007	Identificar de forma clara los elementos que intervienen en el proceso de calendarización de cursos de la DGSCA.
4. Recopilación de Información	Carlos Reyes García Líder del proyecto	Febrero 22 de 2007	Desarrollar las herramientas que servirán para la obtención de la información referente al proceso de calendarización de cursos.
5. Preprocesamiento de los datos	Carlos Reyes García Líder del proyecto	Febrero 26 de 2007	Preparar el conjunto de datos a utilizar a través de la selección, limpieza, integración y transformación de los datos, que sirvan de entrada a un Modelo de Minería de Datos.
6. Modelado	Carlos Reyes García Líder del proyecto	Febrero 26 de 2007	Generación de modelos de Minería de Datos.
	Carlos Reyes García Líder del proyecto	Marzo 26 de 2007	Realizar un análisis estadístico de los datos, representaciones gráficas y generación de algoritmos de minería de datos.

7. Evaluación	Carlos Reyes García Líder del proyecto	Julio 6 de 2007	Verificación de los datos a través del análisis de Minería de Datos realizado y la determinación de la Dirección de Cómputo para la Docencia sobre si la información presentada representa conocimiento nuevo.
8. Despliegue.	Carlos Reyes García Líder del proyecto	Agosto 1 de 2007	Presentación de los resultados obtenidos a la Dirección de Cómputo para la Docencia de la Dirección General de Servicios de Cómputo Académico de la UNAM.

5.1.4.4. Herramientas a utilizar

Las herramientas a utilizar incluyen:

- Equipo de cómputo con configuración de Hardware estándar con el siguiente software precargado:
  - Suite de Microsoft Office 2003 [Word, Excel, Visio, PowerPoint, Project]
  - SQL Server 2000
  - MatLab R13
  - Weka 3.4
  - Cristal Reports XI
  - Adobe Acrobat Profesional
- Acceso a Internet
- Acceso a la Base de Datos de SAEC
- Impresora

5.2. Comprensión de los datos

5.2.1. Compilación inicial de los datos

5.2.1.1. Reporte inicial sobre la adquisición de los datos

La Dirección de Cómputo para la Docencia (DCD) a través de la Coordinación de Producción Académica dirigida por la LI. Juana Figueroa Reséndiz ha proporcionado como fuente de información un respaldo de la base de datos del sistema SAEC[5], cuya finalidad es la explotación de información dentro la misma para llevar a cabo el estudio de Minería de Datos concerniente a este caso de estudio.

[5] Sistema de Administración de Educación Continua.

La base de datos de prueba se encuentra alojada en un servidor de la institución, el RDBMS[6] que se utiliza es MS SQL Server 2000, en donde para tener acceso al mismo se ha provisto al equipo de trabajo de una cuenta de acceso confidencial.

Los datos que se utilizarán corresponden a registros históricos de los cursos impartidos en la DCD, tomando en cuenta un periodo de año y medio.

Para conseguir acceso a la base de datos de prueba de SAEC hay que considerar los elementos propios de una arquitectura cliente – servidor en un RDBMS, tal como se muestra en la figura siguiente.

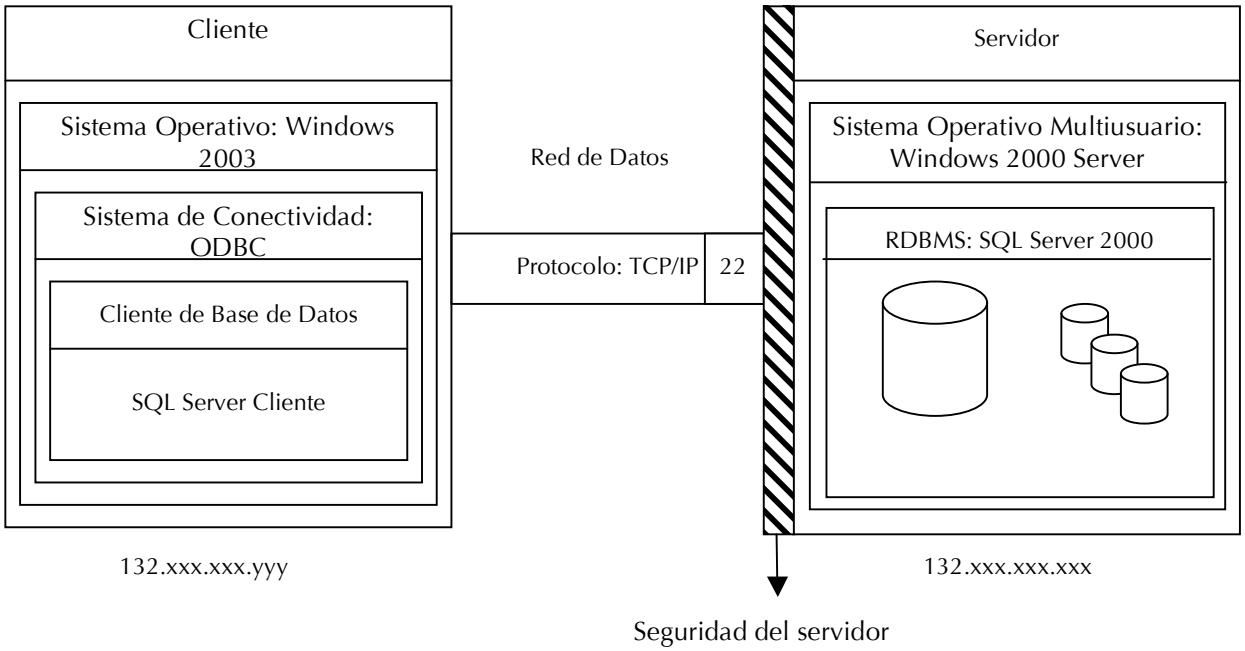

Figura 23. Arquitectura cliente servidor.

5.2.2. Descripción de los datos

5.2.2.1. Reporte sobre la descripción de los datos

Los datos presentes en la base de datos proporcionada la DCD de la DGSCA, UNAM, siguen el modelo de Base de Datos relacional; debido a cuestiones de confidencialidad no es posible publicar el diseño de la base de datos de la institución, sin embargo, para efectos prácticos se presenta un diagrama genérico que refleja una solución a la problemática de almacenar la información generada por concepto de impartición de cursos.

[6] Relational Database Management System (Sistema Manejador de Base de Datos Relacional)

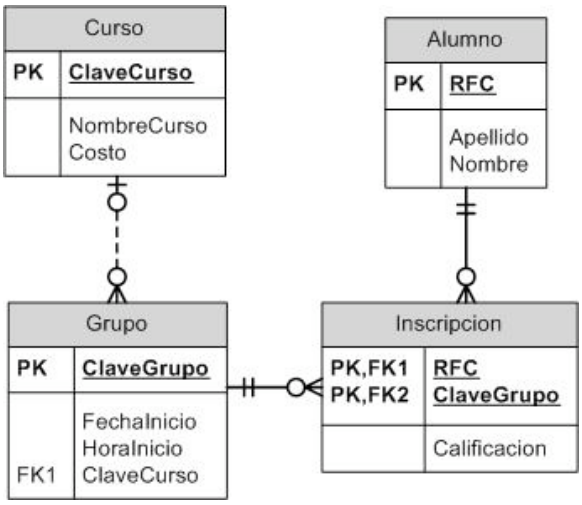

Figura 24. Diagrama Entidad – Relación: Escuela.

Los datos presentes en la base de datos de prueba del SAEC comprenden registros históricos sobre los cursos impartidos en la DCD de la DGSCA, dichos históricos serán suficientes en cantidad para llevar a cabo el estudio de Minería de Datos propuesto; el formato de los datos es adecuado, y respecto de la calidad, podemos decir que la información presente en dicha base de datos no se encuentra homogenizada ya que se ha detectado la existencia de múltiples valores nulos, los cuales tendrán que ser tratados de manera especial.

5.2.3. Exploración de los datos

5.2.3.1. Reporte de la exploración de los datos

Realizada la exploración de los datos se ha determinado que la relación entre las variables es contundente, los datos se encuentran organizados de forma adecuada lo que permitirá generar variables derivadas sin ningún problema, las consultas de selección de información en la base de datos serán vitales para la correcta integración de un conjunto de datos que permitan llevar a cabo el análisis de *clusters*.

En el análisis exploratorio de los datos se ha encontrado que existe una gran propensión sobre la impartición de cursos básicos, los beneficios generados por este tipo de cursos parecen ser prominentes respecto de los de otro tipo, dentro del análisis de *clusters* y haciendo un análisis más profundo, se deberá comprobar esta situación evaluando un conjunto más amplio de variables.

5.2.4. Verificación de los datos

5.2.4.1. Reporte de la calidad de los datos

Una vez analizada la estructura de la base de datos de SAEC y realizada la exploración inicial de los datos se ha determinado que el conjunto de variables a utilizar para el estudio de Minería de Datos se encuentra en un estado permisible, sin embargo se denotan las siguientes observaciones:

- No existe en su totalidad el conjunto de variables idóneas para la realización del análisis de *clusters*. Se deberá generar un conjunto de variables derivadas a partir de la información existente en la base de datos.

- Existencia de valores nulos. Será preciso homogenizar los valores nulos ya que los campos que contienen estos valores son claves para el estudio de Minería de Datos; Homogenizar la información resultará vital para el proyecto, ya que prescindir de los registros que contienen valores nulos afectará de manera contundente el resultado del estudio.
- La confidencialidad de los datos es fundamental por lo que no será posible publicar información que revele la forma en que opera la DCD de la DGSCA, UNAM.
- El acceso a la base de datos de SAEC se ha realizado sin inconvenientes.

5.3. Preparación de los datos

5.3.1. Selección de datos

El caso de aplicación práctico tiene como objetivo evaluar la efectividad de los cursos tomando en cuenta 10 factores que se analizarán multivariadamente con el objeto de reducir la dimensión del problema de 12 a 1 variable que explique la efectividad de los cursos impartidos por la DCD de la DGSCA. Los datos a utilizar en el proyecto comprenden un resumen histórico de los registros contenidos en la base de datos de SAEC, siendo fundamentales las siguientes variables para el estudio:

Nombre del curso	Los campos de resumen que se utilizarán para el estudio, abarcarán información contenida en la base de datos de SAEC para un periodo comprendido entre 1 de enero de 2006 al 15 de junio de 2007
Número de veces impartido	
Alumnos Atendidos	
Alumnos Aprobados	
Alumnos Reprobados	
Cuota	
Honorarios de Profesores	
Gastos Administrativos	
Recuperación Mínima	
Ingresos	
Costo de las Becas	
Monto Neto	
Utilidad	

El listado anterior de variables que será incluido para el estudio de Minería de Datos, tiene la peculiaridad de que se generará a través de diversas operaciones entre los campos encontrados en las tablas de la base de datos de SAEC.

Es preciso incluir tales variables ya que el objetivo del estudio es medir la efectividad de los cursos de cómputo; debido a que la efectividad no es una característica que se pueda medir a través de una sola variable, se realizará un análisis multivariado de datos con las características que definen a un curso, dichas características están definidas en el listado anterior de variables a utilizar en el proyecto.

5.3.2. Limpieza de datos

5.3.2.1. Reporte sobre la limpieza de datos

Tomando como base el reporte de verificación de datos, las actividades realizadas en esta fase incluyeron:

- Eliminación de valores nulos. Dentro del conjunto de datos inicial, proporcionado en la base de datos de prueba SAEC, se encontró un conjunto de registros que contaban con valores nulos entre sus campos, las acciones ejecutadas fueron:
  - Homogenización de valores nulos. Principalmente en los registros referentes a las inscripciones se detecto que no se hallaba la información referente al pago de los alumnos, para cubrir estos valores se tomo en cuenta información de registros con características similares para la homogenización de datos.
  - Actualización de valores en los campos. Se actualizó la información referente a las cuotas de los cursos, con el objeto de realizar un trabajo con cuotas referentes al año en que los cursos fueron impartidos.
- Construcción de variables derivadas. Ésta operación se reporta en la siguiente tarea.

Así pues, los datos al final de esta tarea se encuentran en perfectas condiciones para su evaluación, análisis y procesamiento de fases subsecuentes.

5.3.3. Construcción de datos

5.3.3.1. Atributos derivados

Las variables a utilizar en el estudio de Minería de Datos no se encuentran contenidas en el conjunto de datos dispuesto por la DCD de la DGSCA, por lo que se prosiguió a la generación de los siguientes atributos derivados[7]:

- Número de veces que se impartió el curso
- Alumnos Atendidos
- Alumnos Aprobados
- Alumnos Reprobados
- Recuperación Mínima Total
- Ingresos Totales
- Costo Total de Becas
- Ingresos Netos o Utilidad

[7] Los atributos derivados resumen información para los cursos impartidos en el periodo del 1 de enero de 2006 al 15 de junio de 2007.

Las operaciones realizadas para la generación de los atributos derivados fueron en su totalidad consultas de selección a la base de datos de SAEC, a través del Lenguaje Estructurado de Consultas SQL, y en dónde se realizó agrupamiento de información y resumen de la misma a través de las funciones de agregado provistas en el estándar ANSI SQL 99.

La generación de los atributos derivados se realizó por separado a través de la construcción de vistas[8] actualizables dentro de la base de datos de prueba de SAEC, en dónde será necesaria su posterior integración.

5.3.4. Integración de datos

5.3.4.1. Reporte sobre la integración de datos

Como bien se mencionó anteriormente, para la construcción del conjunto de datos a utilizar en la herramienta de Minería de Datos, se requirió de la generación de atributos derivados, los cuales se encuentran contenidos en vistas actualizables dentro de la base de datos; la idea principal fue la generación de un pequeño Dataware House que previera el conjunto final de datos a utilizar en el proceso de Minería de Datos.

El siguiente esquema, muestra de manera general como fue el proceso de construcción del conjunto de datos a utilizar a través de vistas actualizables dentro de la base de datos.

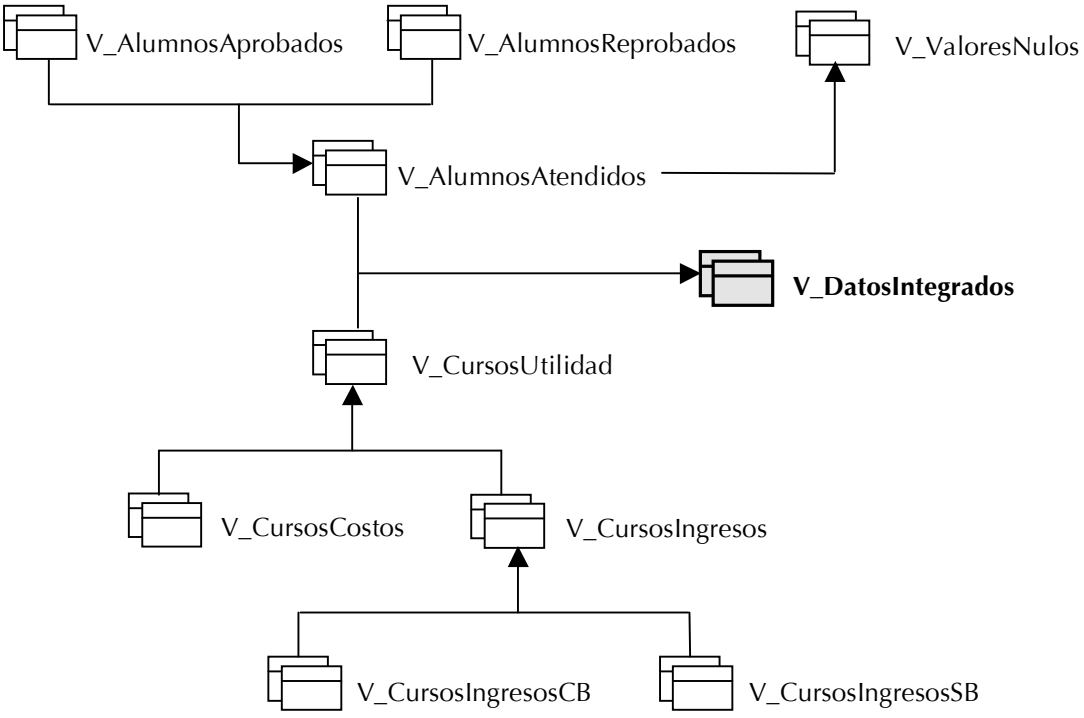

Figura 25. Esquema de vistas para la construcción del DataSet[9].

[8] Vista: Objeto de una base de datos relacional. Query encapsulado en un nombre de objeto.

[9] DataSet o Conjunto de Datos que será utilizado para la ejecución del proceso de Minería de Datos.

Como se muestra en el esquema anterior, la preparación del DataSet consistió en la generación de vistas dentro de la base de datos, las cuales contemplan la construcción de las variables derivadas mencionadas anteriormente; también es observable que se trataron dos conjuntos de vistas principalmente, las referentes a los alumnos y las referentes a los cursos: sus costos e ingresos, para que al final se obtuviera una vista de los datos integrados.

Se construyó también una vista para monitorear valores nulos existentes en los registros de la base de datos, ya que de esa forma se obtiene una lista de los registros a homogenizar dentro de la base de datos; una vez tratados los valores nulos es como se validó el conjunto final de datos integrados.

5.3.5. Formateo de datos

5.3.5.1. Conjunto de datos con formato

Hasta este punto del proyecto se ha determinado que el estudio de Minería de Datos a realizar será la realización de un análisis de clusters, en dónde se obtengan los grupos de cursos que representan un mayor beneficio para la DCD de la DGSCA, UNAM; sin embargo no se ha mencionado como es que se implementará el análisis mencionado, como primer acercamiento señalaré que la implementación del análisis de clusters se hará con la técnica de componentes principales, la cuál se explica más adelante.

La implementación computacional del análisis de componentes principales se realizará con la herramienta MatLab[10] la cuál permite el tratamiento matemático y estadístico de variables numéricas; el manejo de datos de tipo carácter se realiza bajo el concepto de matriz de datos en donde los datos que serán procesados contengan una longitud igual de caracteres.

Por consiguiente la entrada de datos al software matemático MatLab tiene el siguiente requisito:

- Las cadenas de datos que se ingresen a MatLab deberán contar con la misma longitud, ejemplo: la cadena de datos “Computadora” tiene una longitud de 11 caracteres, mientras que la cadena de datos “Teclado” tiene una longitud de 7 caracteres, para el correcto funcionamiento de estos datos en software MatLab se deberán homogenizar todas las cadenas de datos a utilizar ajustando el tamaño de las mismas a la de mayor longitud; por lo tanto la cadena de datos “Teclado” deberá tener una longitud de 11 caracteres complementando ésta con espacios en blanco: “Computadora”, “Teclado ”, así ambas cadenas tendrán una longitud igual.

Dentro del conjunto de datos a utilizar en el proyecto se ha detectado que se requerirá homogenizar el contenido de la variable “Nombre de curso”, ya que es de tipo cadena y todos los valores que adopten dicha variable deberán contener la misma longitud.

[10] MatLab es un software computacional para el procesamiento de datos, de aplicación matemática y para la generación de gráficos especializados sobre variables numéricas.

3.3.6. Resultados de la fase

Como resultado de la fase "Preparación de los datos", se entrega el conjunto de datos a utilizar en la fase de "Modelado". El conjunto de datos a utilizar contempla información sobre los siguientes rubros:

Clave	Atributo	Descripción
V_1	Nombre del curso	Nombre del curso que imparte la DCD de la DGSCA, UNAM.
V_2	Número de veces impartido	Número de veces que se impartió el curso en el periodo[11].
V_3	Alumnos Atendidos	Número de alumnos atendidos para el curso determinado en el periodo.
V_4	Alumnos Aprobados	Número de alumnos que aprobaron el curso.
V_5	Alumnos Reprobados	Número de alumnos que reprobaron el curso.
V_6	Cuota	Promedio de la cuota o costo del curso durante el periodo señalado.
V_7	Honorarios de Profesores	Suma del costo por concepto de pago de honorario a profesores de un curso en el periodo determinado.
V_8	Gastos Administrativos	Suma de los gastos administrativos generados por concepto de un curso en el periodo determinado.
V_9	Recuperación Mínima	Monto total esperado como recuperación mínima al final del periodo determinado.
V_{10}	Ingresos	Ingresos recabados por concepto de impartición de un curso al cabo del periodo señalado.
V_{11}	Costo de las Becas	Costo total del otorgamiento de becas al cabo del periodo señalado.
V_{12}	Monto Neto	Monto – Descuentos
V_{13}	Ingresos Netos	Ingresos – (Honorarios de Profesores + Gastos Administrativos + Costo de Becas)

[11] Periodo comprendido entre 1 de enero de 2006 al 15 de junio de 2007.

Por cuestiones de confidencialidad no es posible la publicación del DataSet, sin embargo, para efectos prácticos piense en una matriz de datos como se muestra a continuación[12]:

V_1: Curso	V_2	V_3	V_4	V_5	…	V_{11}	V_{12}
Access	25	350	340	10		$1,500	$8,500
Word	50	600	545	55		$3,500	$9,600
Excel	14	580	563	17		$6,280	$3,500
Java	65	640	600	40		$2,350	$9,500
.NET	90	720	712	8		$1,850	$1,200
Internet	150	360	360	0		$890	$6,700

5.4. Modelado

5.4.1. Elección del modelo

5.4.1.1. Definición del modelo a utilizar en el proyecto

Como bien ya se ha mencionado, el objetivo del proyecto en términos de Minería de Datos consiste en evaluar la efectividad de los cursos de cómputo a través de un Análisis de Componentes Principales tomando en cuenta 12 factores que se analizarán multivariadamente con el objeto de reducir la dimensión del problema de 12 a 1 variable que explique la efectividad de los cursos impartidos por la DGSCA.

Una vez evaluados los datos y preprocesados, se ha determinado la utilización del Análisis de Componentes Principales (ACP) como modelo de Minería de Datos a utilizar en el proyecto; se realizó esta elección, ya que, ACP es una técnica de análisis multivariado de datos que tiene un gran rendimiento sobre variables presentadas a escalas diversas, es decir ACP es menos susceptible de errores cuando se evalúan variables a escalas totalmente opuestas, siendo este el caso encontrado en las diversas variables que caracterizan a un curso de cómputo. Por ejemplo suponga que se desea hacer un análisis multivariado de datos con la variable 'x' y la variable 'z', cuyos dominios son:

- *'x'*={1,2,3,4,5}
- *'z'*={15000, 250000, 1500000, 65000, 150000}

Al momento de ponderar el peso específico de cada variable para un caso particular, nos encontramos con una diferencia enorme, ya que en principio podríamos decir que *'z'* tiene mayor peso respecto de *'x'*, algo que puede ser cierto pero que no podemos afirmar a simple vista. Cuando se utiliza ACP, a diferencia de otros modelos de Minería de Datos, la primera tarea a realizar consiste en eliminar la diferencia escalar entre las variables, lo que la hace una técnica con gran rendimiento y con menor susceptibilidad a cometer errores.

[12] Datos de ejemplo que no reflejan el modelo de negocio que sigue la DCD de la DSGCA, UNAM.

5.4.2. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) pertenece a un grupo de técnicas estadísticas multivariantes, eminentemente descriptivas.

"El ACP permite reducir la dimensionalidad de los datos, transformando el conjunto de ***p*** variables originales en otro conjunto de ***q*** variables incorrelacionadas (***q*** <= ***p***) llamadas componentes principales. Las ***p*** variables son medidas sobre cada uno de los ***n*** individuos, obteniéndose una matriz de datos de orden ***np*** (***p*** < ***n***)." (González, Díaz, Torres y Garnica 1994: 55 - 57)

"El Análisis de Componentes Principales (ACP) es una técnica de descripción estadística para la visualización aproximada (pero en cierto modo optimal) de la información contenida en una tabla de datos: descripción simultánea de la asociación existente entre variables y similitud de individuos. Es también una técnica de reducción de la dimensionalidad de un conjunto de variables continuas, utilizable como paso intermedio con vista a análisis ulteriores." (Aluja 1999: 15)

La tabla de datos que se utilizará para el caso de estudio práctico, consiste de un conjunto de doce variables que caracterizan a los cursos de cómputo que imparte la DCD de la DGSCA. Las dimensiones de la tabla hacen imposible observar mediante la simple inspección, cuáles son los individuos (cursos) similares, ni que variables miden cosas parecidas en los distintos individuos.

El patrón de asociación entre variables y la configuración de las similitudes entre individuos, que constituye la síntesis de información contenida en la tabla, permanece escondida, siendo el ACP el encargado de revelar esta información.

Así pues, podemos decir que la realización de un ACP consiste en primer lugar en hacer una explicación sobre la correlación existente entre las variables y en segundo término reducir la dimensionalidad de las variables que caracterizan a un individuo con el objeto principal de hacer un análisis de las distancias (diferencias entre individuos) que de pauta a la generación de *clusters*.

Para dar una explicación sobre la correlación entre variables, existe la opción de usar la matriz de correlaciones[13] o bien, la matriz de covarianzas[14]. En la primera opción se le está dando la misma importancia a todas y a cada una de las variables; esto puede ser conveniente cuando el investigador considera que todas las variables son igualmente relevantes. La segunda opción se puede utilizar cuando todas las variables tengan las mismas unidades de medida y además, cuando el investigador juzga conveniente destacar cada una de las variables en función de su grado de variabilidad.

Para lograr la reducción en la dimensionalidad del problema y la generación de *clusters* (grupos), el ACP toma como base una aproximación geométrica bastante simple e intuitiva, que está en la base y es el fundamento de todos los análisis factoriales exploratorios, y que consiste en asociar a la tabla de datos dos nubes de puntos: la nube de puntos fila y la nube de puntos columna.

5.4.2.1. La nube de puntos fila

Por un lado, podemos considerar cada fila de la tabla de datos, en nuestro caso un curso de cómputo, como un punto de coordenadas iguales a los distintos valores tomados por las variables activas, esto es, en nuestro caso las 12 características de los cursos.

[13] Matriz de datos que explica la relación entre variables dada en términos de varianza.

[14] Matriz de datos que explica el grado de variabilidad de una variable comparada con otras, en términos de covarianza.

Sí solamente tuviéramos registradas 3 características de los cursos de cómputo, los valores tomados por cada individuo lo situarían en un espacio tridimensional, definiendo una nube de 55 puntos (cursos).

Ahora bien, si 3 características o variables sitúan a un curso en un espacio tridimensional, imagine que pasa cuando se evalúan 12 características para un curso.

En general se obtiene una nube de ***n*** individuos (tantos como filas de la tabla de datos) en un espacio de ***p*** dimensiones (tantas como variables se tengan). Llamamos a esta nube formada, la nube de puntos fila o nube de los individuos.

En esta nube es obvio que dos puntos aparezcan próximos uno del otro, estén reflejando dos cursos de cómputo con características similares entre las 12 observancias hechas. Por el contrario, dos puntos que aparezcan alejados uno de otro están indicando dos cursos de cómputo con características diferentes.

Para medir la noción de proximidad entre puntos-fila (cursos en nuestro caso), debemos definir una distancia entre cursos de cómputo.

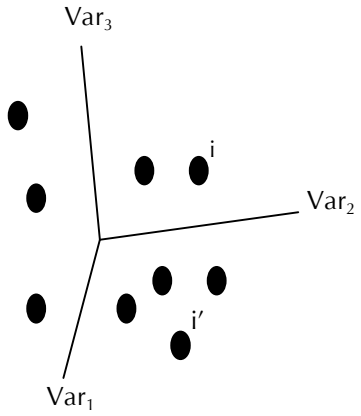

Figura 26. Nube de puntos fila

La medida de distancia más intuitiva es la distancia euclidiana entre puntos.

$$d^2(i,i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Figura 27. Fórmula de la distancia Euclidiana

El problema para poder visualizar estas distancias entre puntos, reside en la elevada dimensionalidad de la nube de puntos-fila (para el caso de estudio 12 dimensiones) que la hace opaca a la comprensión humana.

5.4.2.2. La nube de puntos columna

De forma análoga, podemos presentar las ***p*** columnas de nuestra tabla como ***p*** puntos de un espacio de ***n*** dimensiones (una para cada individuo). Las ***n*** coordenadas de punto columna, vienen definidas por los ***n*** valores de la variable correspondiente en la tabla de datos.

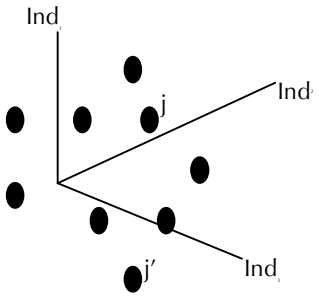

Figura 28. Nube de puntos columna o puntos variable

El interés de esta nube de puntos variable, se encuentra centrada en que es una representación de las asociaciones existentes entre ellas. Cada variable mide una característica observada sobre los cursos de cómputo y, por tanto, podemos ver que variables miden cosas parecidas en los cursos. Será interesante, pues, definir una distancia entre puntos-variable que expresa la densidad y la naturaleza de la asociación entre las variables. Dos puntos-variable próximos indicarán dos variables que toman valores relacionados en el conjunto de cursos: sabiendo el valor de una, podremos prever el valor en la otra variable.

Una medida muy usual para medir la asociación entre las variables es el coeficiente de correlación lineal entre ellas. Si utilizamos este coeficiente como base de la distancia entre variables, entonces la visualización de la nube puntos-variable se convierte en la visualización de la matriz de correlación entre variables.

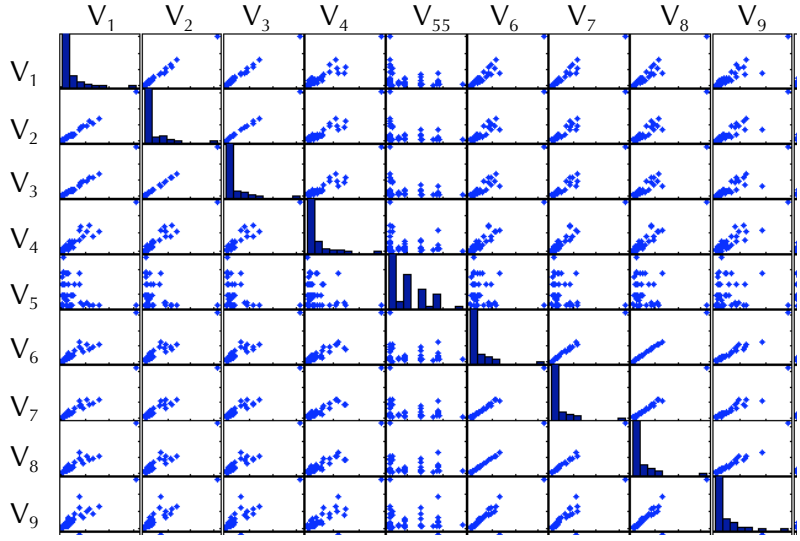

Figura 29. Matriz de correlación lineal entre variables

5.4.2.3. Resultados del ACP

Las ***p*** nuevas variables generadas a partir de la reducción de dimensionalidad (componentes principales) son obtenidas como combinaciones lineales de las variables originales. Los componentes se ordenan en función del porcentaje de varianza explicada. En este sentido, el primer componente será el más importante por ser el que explica mayor porcentaje de la varianza de los datos. Queda a criterio del investigador decidir cuántos componentes se elegirán en el estudio.

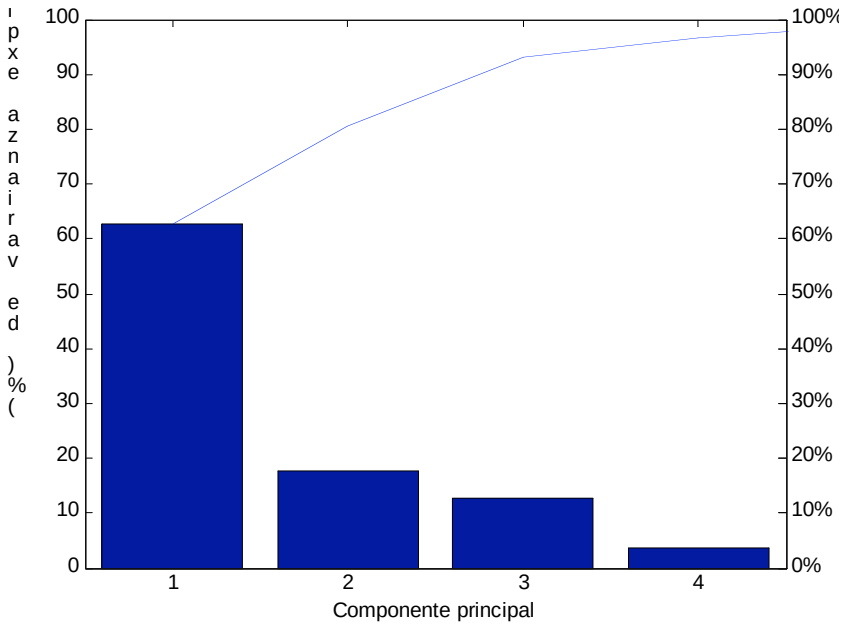

Figura 30. Componentes principales y su respectivo porcentaje de varianza.

El ACP se realiza en el espacio de las variables y, en forma dual, en el espacio de los individuos (nube de puntos fila y nube de puntos columna). Se acostumbra a representar gráficamente los puntos-variables y los puntos-individuos tomando como ejes de coordenadas los componentes. A veces, puede facilitar la interpretación de los resultados, el observar la similar ubicación de los puntos en los planos respectivos. Aunque el plano de puntos-variables no se superpone al plano de puntos-individuos, es de gran utilidad "interpretar" la cercanía de un grupo de puntos-individuos, a ciertas variables.

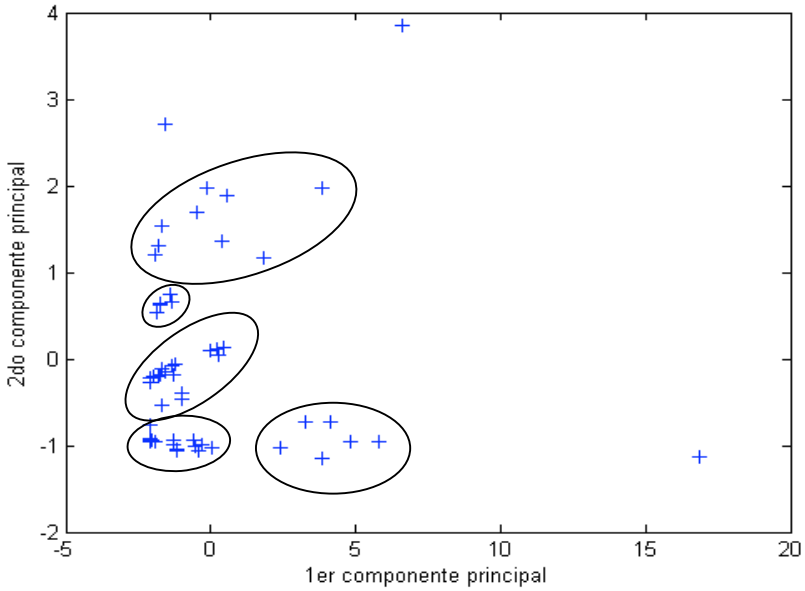

Figura 31. cercanía de un grupo de puntos-individuos, a ciertas variables.

5.4.2.4. Criterios para seleccionar el número de componentes

Uno de los objetivos básicos del ACP es reducir la dimensionalidad de los datos. Lo ideal sería seleccionar el primero o los dos primeros componentes, ya que de esta forma se puede analizar en una recta o en un plano las posibles interrelaciones existentes entre los puntos variables y los puntos individuos.

Los criterios que más se utilizan para decidir el número de componentes a seleccionar son:

- Gráfico de autovalores: consiste en representar el porcentaje de variación explicada contra el número del componente. En el eje de las ordenadas se registra el porcentaje de variación explicada. En el eje de las abscisas se coloca el número del componente según su orden de importancia de acuerdo a la variación explicada. Por lo general, los puntos del gráfico presentan una figura similar al perfil de una bota. Al analizar este gráfico se busca el punto de quiebre, donde el cambio de la pendiente se hace mayor, y la abscisa correspondiente a este punto indica el número de componentes a retener.
- Promedio de autovalores: se calcula el promedio de todos los autovalores y se eliminan aquellos autovalores que están por debajo de este promedio. Si se trabaja con la matriz de correlaciones, este promedio es uno. Mediante este criterio se tiende a retener menos componentes que en el criterio anterior y aunque es más objetivo, puede considerarse menos flexible ya que en el primer caso el investigador puede elegir el número de componente haciendo uso de su experiencia personal.

Aparte de los criterios señalados, es importante atender otros aspectos relacionados con la naturaleza de la investigación. Por ejemplo, un componente con muy poca contribución puede estar altamente

correlacionado con alguna variable importante en la investigación, entonces no sería conveniente desincorporar este componente en el análisis final.

5.4.3. Diseño de pruebas

5.4.3.1. Documentación de pruebas

La validez del modelo de Minería de Datos se realizará a través de la representación que se haga de los componentes principales generados, de forma tal que se obtenga el número preciso de componentes principales que expliquen el fenómeno de estudio con un nivel de confianza igual o mayor al 90%.

Una componente principal explica un fenómeno en porcentajes de varianza, para probar el modelo de Minería de Datos se deberán generar las componentes principales necesarias hasta alcanzar un nivel de confianza del 90%. El gráfico siguiente muestra un conjunto de componentes principales que explican un fenómeno, para validar el modelo se deberán reunir tres componentes principales, ya que así, se alcanza un nivel de confianza superior al 90%.

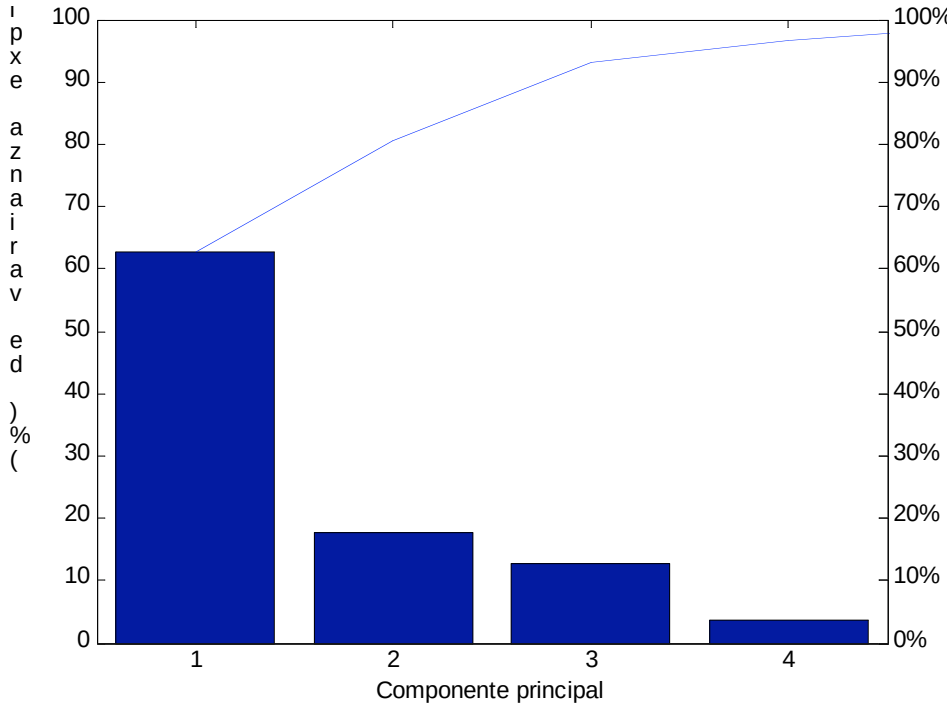

Figura 32. Tres componentes principales explican el fenómeno en más del 90%.

Por tanto, el diseño de pruebas consiste en generar un gráfico con los componentes principales obtenidos y evaluar si cumplen con el nivel de confianza antes señalado, de no ser así, la implementación del modelo deberá llevarse acabo nuevamente, ajustando los elementos que permitan obtener los componentes principales requeridos.

El conjunto de datos a utilizar no se dividirá para entrenar el modelo, estos pasarán íntegros al modelo, lo que probará la calidad del modelo será el porcentaje de varianza en un 90% a través de *n* componentes principales.

5.4.4. Construcción del modelo

5.4.4.1. Descripción del modelo

La implementación del modelo de Minería de Datos a utilizar se realizó en el software matemático y estadístico MatLab. Para la implementación del Análisis de Componentes Principales se utilizó el siguiente algoritmo:

Entradas:

A. DataSet del fenómeno a evaluar en donde a partir de *n* variables o dimensiones que definen un objeto, se reducirán para obtener una variable global que resuma las características del problema y así generar grupos (*clusters*) de los cursos impartidos por la DCD de la DGSCA, UNAM.

V1: Curso	V2	V3	V4	V5	…	V11	V12
Access	25	350	340	10		$1,500	$8,500
Word	50	600	545	55		$3,500	$9,600
Excel	14	580	563	17		$6,280	$3,500
Java	65	640	600	40		$2,350	$9,500
.NET	90	720	712	8		$1,850	$1,200
Internet	150	360	360	0		$890	$6,700

Proceso:

1. Generar un gráfico de caja y bigote con los datos crudos.
2. Estandarizar los datos contenidos en DataSet de cursos.

   Objetivo: Analizar los datos en una sola escala. La estandarización de los datos se consigue dividiendo cada uno de los valores del DataSet entre la desviación estándar de los mismos para cada una de las variables.

3. Generar un gráfico de caja y bigote con los datos estandarizados.

   Objetivo: Hacer un análisis de la información proporcionada por las variables. Resaltar cursos que sobresalen en diversos ámbitos.

4. Normalizar los datos ya estandarizados.

   Objetivo: Reducir el espectro escalar de los datos para analizarlos en un rango común.

5. Generar un gráfico de caja y bigote con los datos normalizados.

   Objetivo: Evaluar las características de los cursos para denotar aquellos que resaltan por arriba o por debajo de la media de valores.

6. Obtener la distancia Euclideana entre los valores de datos normalizados y ligar los datos.

   Objetivo: Dividir un conjunto de objetos en grupos (cluster en inglés) de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y los de los objetos de

clusters diferentes sean distintos (aislamiento externo del grupo). Esto es posible ya que, para obtener la distancia Euclideana se toman los datos normalizados y se resume el universo de variables estudiadas a una sola dando una explicación global del fenómeno.

7. Generación de un dendograma.

   Objetivo: El dendograma muestra el resultado de la generación de clusters, ya que es un gráfico que muestra las subclases obtenidas.

8. Obtener la matriz de correspondencia de correlación de matrices.

   Objetivo: La matriz de correspondencia de matrices permite observar la determinación que existe en una variable a partir de las demás variables estudiadas. Es decir a través de ésta matriz podemos por ejemplo, afirmar que la variable 'x' esta determinada por 'y'; siempre que sea 'y', será 'x'.

9. Obtener las componentes principales que den explicación al fenómeno estudiado.

   Objetivo: Probar y validar el modelo generado a través de la generación de componentes principales que dan explicación al fenómeno.

10. Obtener un índice para listar los cursos de cómputo analizados de acuerdo a su efectividad, tomando el análisis previo en dónde se redujo la dimensionalidad del problema de *n* a 1 variable.

    Objetivo: Mostrar el conjunto de cursos estudiados ordenándolos de acuerdo a la efectividad encontrada, dicha efectividad resume la calidad de un curso en términos económicos y académicos.

Salidas:

A. Durante el proceso se generan diversos gráficos, los cuales ya han sido mencionados, dichos gráficos son salidas de información que deberán ser analizados por especialistas de Minería de Datos para la correcta interpretación de información. Dichos gráficos son:

   - Gráfico de caja y bigote con datos crudos.
   - Gráfico de caja y bigote con datos estandarizados.
   - Gráfico de caja y bigote con datos normalizados.
   - Dendograma que muestra las subclases contenidas para los cursos de cómputo analizados.
   - Gráfico de componentes principales.

B. índice de cursos ordenado de acuerdo a la efectividad de los mismos.

Como resultado, ahora se tiene una librería de MatLab la cuál es ya genérica para la implementación del Análisis de Componentes Principales en situaciones futuras.

Para emplear la librería antes mencionada bastará enviar como parámetros el DataSet generado en fases anteriores, como resultado se arrojará un listado de los cursos de cómputo que imparte la DCD de la DGSCA, UNAM ordenados de acuerdo a su nivel de eficiencia, también se arrojará un conjunto de gráficos que deberán ser interpretados por especialistas en datos. Todos los elementos antes mencionados conforman el conocimiento generado en el proceso de Minería de Datos.

5.4.5. Evaluación del modelo

5.4.5.1. Reporte sobre la evaluación del modelo

La implementación del modelo se llevó a cabo con éxito, se han corrido diversos conjuntos de datos, arrojando resultados satisfactorios. El modelo se encuentra listo para el análisis sobre los datos referentes al fenómeno de estudio en cuestión.

5.4.5.2. Revisión de los parámetros

Los parámetros requeridos por el modelo se encontraron en correctas condiciones, es decir el DataSet se encuentra en forma para su utilización dentro del modelo de Componentes Principales ya generado.

La restricción con que se contaba era la siguiente:

- Las cadenas de datos que se ingresen a MatLab deberán contar con la misma longitud.

5.5. Evaluación

5.5.1. Evaluación de resultados por los especialistas en Minería de Datos

Una vez realizada la implementación del Análisis de Componente Principales en una librería de MatlLab, se procedió a su ejecución en dónde se obtuvieron los siguientes resultados.

Como primer paso se envía como parámetro a MatLab el DataSet generado en la fase de *preparación de los datos*, el cual contiene información sobre los cursos que imparte la DCD de la DGSCA, UNAM.

Variables[15]**:**

- Nombre del curso
- Número de veces impartido
- Alumnos Atendidos
- Alumnos Aprobados
- Alumnos Reprobados
- Cuota
- Honorarios de Profesores
- Gastos Administrativos
- Recuperación Mínima
- Ingresos
- Costo de las Becas
- Monto Neto
- Ingresos Netos o Utilidad

[15] Información referente al periodo comprendido del 1 de enero de 2006 al 15 de junio de 2007.

V1: Curso	V2	V3	V4	V5	…	V11	V12
Access	25	350	340	10		$1,500	$8,500
Word	50	600	545	55		$3,500	$9,600
Excel	14	580	563	17		$6,280	$3,500
Java	65	640	600	40		$2,350	$9,500
.NET	90	720	712	8		$1,850	$1,200
Internet	150	360	360	0		$890	$6,700

Una vez cargados los datos en el software MatLab, se ejecuta la librería implementada para el análisis de componentes principales.

1. Una vez cargados los datos, la primera tarea que se realizará es la generación de un grafico de caja y bigote con los datos crudos[16]; un tipo de gráfico estadístico para resumir información utilizando 5 medidas estadísticas: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo.

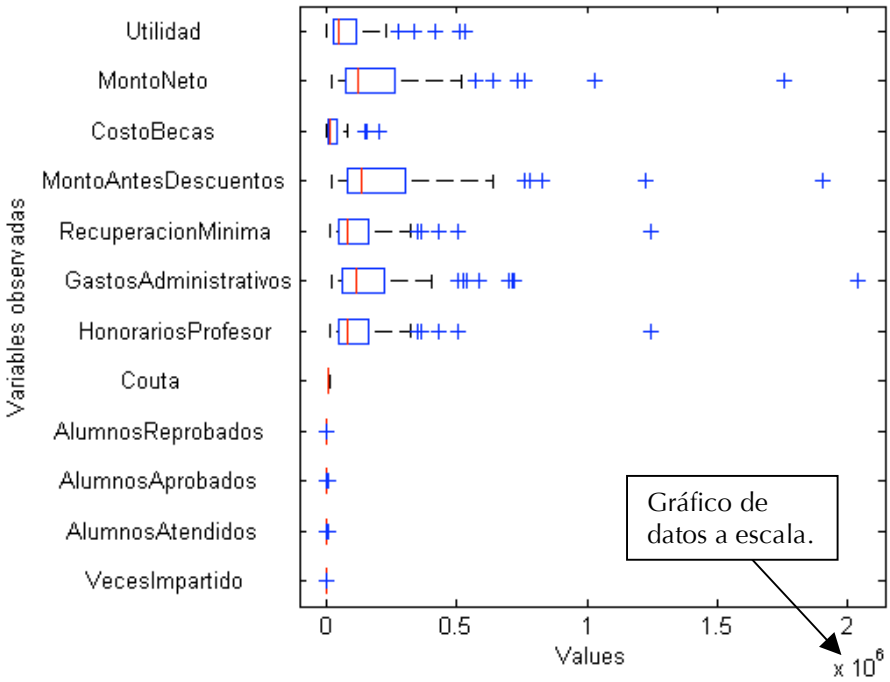

Figura 33. Gráfico de caja y bigote utilizando datos crudos.

La interpretación que generalmente se hace para un gráfico de caja y bigote, consiste en observar la forma en que están distribuidos los valores para cada una de las variables graficadas, siendo posible observar el valor mínimo (cruz azul a la extrema izquierda de la caja), el primer cuartil, la mediana (línea roja), el tercer cuartil y el valor máximo (cruz azul a la extrema derecha de la caja).

[16] Datos que no tienen más que el procesamiento para su generación, los valores se encuentran en las unidades y espectros reales.

El gráfico anterior resulta de poca utilidad debido a que está presentado a escala ($_x10^6$), siendo imposible la visualización e interpretación correcta del mismo. Por lo tanto se puede concluir que los datos no se aprecian bajo un mismo rango de valores y por tanto no es posible su correcta interpretación.

2. Generación de un gráfico de caja y bigote con datos estandarizados.

La estandarización de los datos es el proceso a través del cual se elimina la escala existente en un conjunto de valores; para el caso de estudio en cuestión, por ejemplo, la variable "Veces impartido" podría contener un rango de valores de 1 hasta 50, mientras que la variable "Utilidad" puede contener valores en el rango de $1,000 a $10,000; así pues, la evaluación de ambas variables resulta difícil ya que su espectro de valores es distinto; para ajustar los valores se requiere de una estandarización, que no es más que la acción de dividir cada uno de los valores de una variable entre la desviación estándar[17] de la misma.

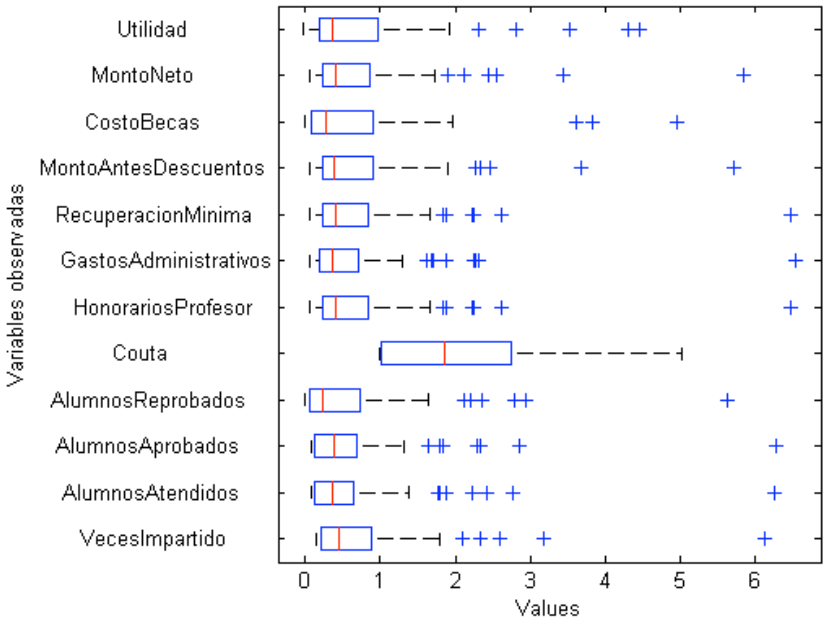

Figura 34. Gráfico de caja y bigote utilizando datos estandarizados.

A diferencia del gráfico anterior, en este, resulta sencillo hacer una evaluación del comportamiento de las variables utilizadas, el gráfico ya no muestra valores a escala, la estandarización de los datos sirvió para realizar una correcta observación de los datos contenidos en las variables.

Debido a que se busca realizar un análisis multivariante de datos, es que se presenta este gráfico de caja y bigote para cada variable, así es posible visualizar la distribución de la información contenida en el DataSet.

Las cruces azules del gráfico, representan valores por encima de la media de datos, es decir, información sobre cursos que sobresalen. Para la variable *Utilidad* por ejemplo, podemos observar un conjunto de cinco cruces a la extrema derecha, las cuales representan valores de cinco cursos de cómputo que proporcionan una mayor utilidad a la institución que imparte los cursos de cómputo.

[17] Medida de dispersión utilizada en la estadística descriptiva que informa de la media de distancias que tienen los datos respecto de su media aritmética.

A su vez, si observamos el gráfico de caja y bigote para la variable *VecesImpartido,* encontramos un conjunto de cinco cruces a la extrema derecha lo que significa que existe un conjunto de cinco cursos que se impartieron en un mayor número de ocasiones, misma situación posiblemente refleje el hecho de que la variable *Utilidad* esté determinada por el número de veces que se imparte un curso, cuestión que resulta ciertamente lógica.

Para los gráficos de caja y bigote de *RecuperaciónMínima*, *GastosAdministrativos* y *HonorariosProfesor* existe una cruz a extrema derecha que refleja el hecho de que un curso tiene costos por encima de la media y en este sentido podemos decir que los costos altos darán como resultado una menor utilidad.

El conjunto de variables con las cuales determinaremos la efectividad de un curso, pueden dividirse en dos rubros, variables que reflejan efectividad positiva (*Ingresos, VecesImpartido, AlumnosAtendidos, etc.*) y variables que reflejan reflejan efectividad negativa o mejor dicho efectividad menor y que restan valor a los cursos que se imparten en la institución, dichas variables son *RecuperaciónMínima, GastosAdministrativos, HonorariosProfesor, etc.*

Tomando en cuenta que se cuenta con dos rubros de variables en donde un grupo de estas suman efectividad y otras restan efectividad a los cursos de cómputo es como se realizará un análisis multivariante en donde se pondere de manera adecuada la participación de cada variable en la efectividad de un curso de cómputo.

$\chi_{12} \rightarrow \chi_1$ en dónde χ representa la efectividad de un curso de cómputo.

Donde $\chi 12$ refleja doce variables que aportan a la efectividad de un curso de cómputo en diversos ámbitos, así, a través del Análisis de Componentes Principales se reflejará en un solo término ($\chi 1$) la efectividad de un curso de cómputo. Hecho esto es como se podrán generar los *clusters*, agrupando los cursos que tengan una efectividad similar tomando en cuenta una sola variable $\chi 1$.

3. Los datos estandarizados son de gran utilidad para ver el conjunto de datos en un mismo espectro de valores, sin embargo para optimizar la visualización de las variables se requiere llevar a cabo una tarea más, la normalización de los datos, la cual tiene por objeto eliminar el ruido encontrado en los valores estándar y a su vez hacer un ajuste mayor respecto al dominio de los valores estudiados para colocarlos en un espectro menor.

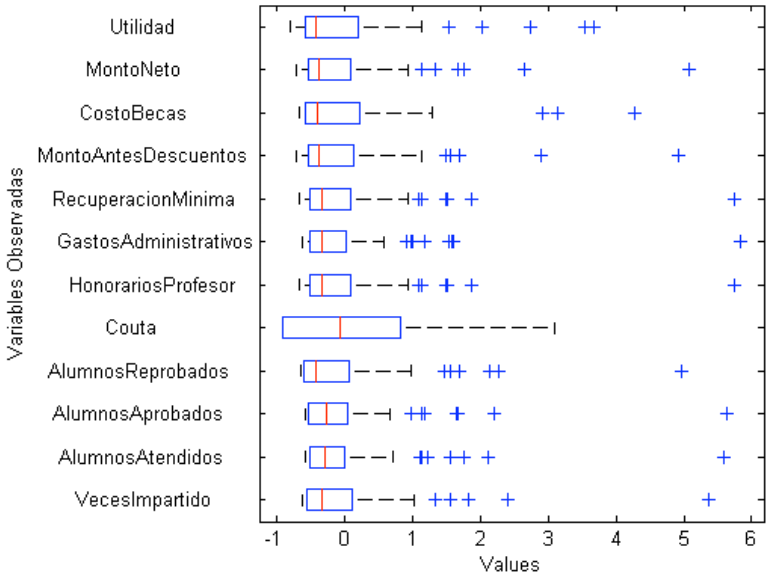

Figura 35. Gráfico de caja y bigote utilizando datos normalizados.

La normalización de los datos y la generación de un gráfico de caja y bigote permiten hacer una observancia de las variables en un sentido óptimo, teniendo la posibilidad de apreciar en un mismo espectro cual es la dispersión de los datos.

El Análisis de Componentes Principales se realizará con los datos estandarizados ya que estos proporcionan un conjunto de valores dentro de un rango similar para cada variable, haciendo posible la comparación entre variables.

Datos crudos:

V1: Curso	V2	V3	V4	V5	…	V11	V12
Access	25	350	340	10		$1,500	$8,500
Word	50	600	545	55		$3,500	$9,600
Excel	14	580	563	17		$6,280	$3,500
Java	65	640	600	40		$2,350	$9,500
.NET	90	720	712	8		$1,850	$1,200
Internet	150	360	360	0		$890	$6,700

Datos estandarizados y normalizados:

V1: Curso	V2	V3	V4	V5	…
Access	0.14615363	0.07705817	0.08071577	0.04810589	
Word	0.14615363	0.07252533	0.07567104	0.04810589	
Excel	0.20461508	0.190379	0.18665522	0.24052943	
Java	0.26307653	0.13145216	0.12611839	0.19242355	
.NET	0.90615248	0.63459665	0.60032356	1.01022362	
Internet	0.17538435	0.13145216	0.14629734	0	

4. Continuando con el proceso referente al Análisis de Componente Principales, el siguiente paso consta de la obtención de la Distancia Euclidiana entre los datos.

   La obtención de la Distancia Euclidiana[18] permitirá realizar la reducción del universo de variables, es decir la forma en que el universo de variables puede resumirse en un solo elemento

   $\chi_{12} \rightarrow \chi_1$ en dónde χ representa la efectividad de un curso de cómputo.

[18] Índice cuantitativo que mide la separación existente entre dos unidades de observación según los valores que ellas posean en un conjunto de variables. (Mendenhall 1998).

La Distancia Euclidiana se obtuvo entra las variables que definen a un curso de cómputo, para el curso de Access por ejemplo, se obtuvo la Distancia Euclidiana entre los siguientes elementos {0.14615363, 0.07705817, 0.08071577, 0.04810589}

La Distancia Euclidiana permitirá ligar los datos para la generación de *clusters*, es decir, será el factor que permita la generación de grupos de cursos de cómputo evaluados de acuerdo a su efectividad.

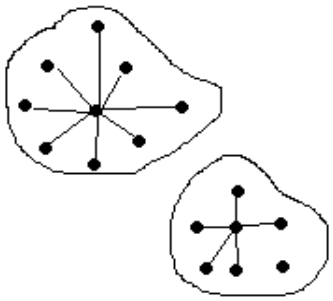

Figura 36. Generación de clusters.

La Distancia Euclidiana será el elemento que permitirá ligar los datos para la generación de grupos. En la figura anterior cada punto es el reflejo de un curso de cómputo, en donde se agrupan aquellos con características similares (efectividad), el factor que determina que elemento pertenece a cada grupo será la Distancia Euclidiana.

5. Una vez que se ha reducido el universo de variables y que se ha generado los agrupamientos pertinentes, es como se genera el siguiente dendograma[19]

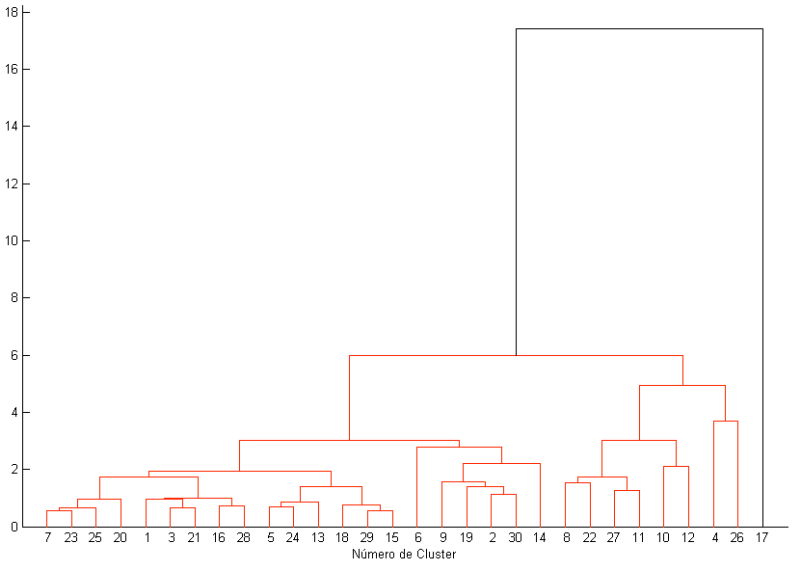

Figura 37. Dendograma, reflejo de los clusters encontrados.

El dendograma generado muestra el número de clusters encontrados en el caso de estudio, cada uno de los clusters contiene la información sobre cursos con efectividad similar; a pesar de que el gráfico muestra un gran número de clusters, es posible ir niveles hacia arriba para agrupar un mayor número de cursos de cómputo.

En la figura siguiente se puede observar la existencia de 5 grupos de cursos bien definidos, de izquierda a derecha se representa la efectividad de cada grupo de cursos, siendo los *clusters* 4 y 26 los que mayor efectividad tienen al ser impartidos por la DCD de la DGSCA, UNAM. El *cluster* 17, resulta un caso muy especial, ya que dicho *cluster* contiene el conjunto de cursos con la mayor efectividad para la institución.

[19] Gráfico que representa el conocimiento obtenido dada una generación de grupos.

Un *cluster* agrupa uno o más cursos de acuerdo al espacio en el cual están dibujados los datos, más adelante se presentará un gráfico que explique esta situación.

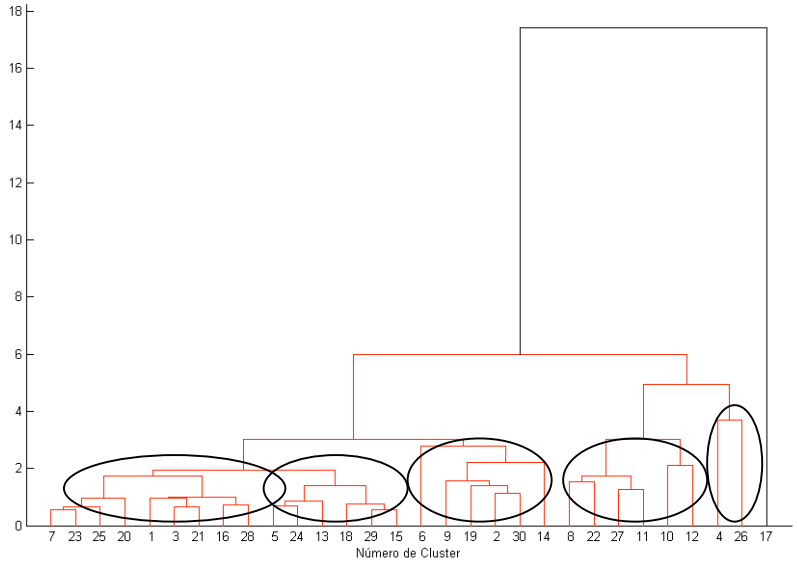

Figura 38. Cinco clusters principales y un caso especial.

En total se han generado 30 grupos de cursos diferentes, recordando que el DataSet inicial cuanta con información de 55 cursos de cómputo. Esto es poco efectivo y por ello se ilustran cinco grupos principales de cursos de acuerdo a la figura anterior, recordando que el cluster 17 es un caso particular.

Por motivos de confidencialidad resulta imposible la publicación de los nombres de los cursos que pertenecen a cada grupo, para ejemplificar esta situación se ha distorsionado la información siguiente:

Suponga que el *cluster* 17 esta integrado por los cursos de Java, Visual Basic y Word, ello significa que estos tres cursos son los de mayor efectividad para la institución y que por tanto su calendarización deberá llevarse a cabo de forma más continua que el resto de los cursos.

El *cluster* número 7 por el contrario, integrado por los cursos Access e Internet representa a cursos con poca efectividad para la institución, lo que obliga a tomar en consideración un conjunto de medidas para elevar los niveles de eficiencia de dichos cursos.

6. Con el objeto de dar un mayor sentido a la explicación del fenómeno en cuestión, se genera la siguiente matriz de correspondencia de variables.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7
V_1	1.00						
V_2	0.99	1.00					
V_3	0.99	1.00	1.00				
V_4	0.94	0.95	0.94	1.00			
V_5	-0.26	-0.26	-0.27	-0.21	1.00		
V_6	0.96	0.96	0.95	0.94	-0.08	1.00	
V_7	0.97	0.97	0.97	0.94	-0.13	1.00	1.00

Figura 39. Matriz de correspondencia.

La matriz de correspondencia de variables, se genera a través de la aportación de varianza en los datos, la varianza en los datos aporta información de cada una de las variables estudiadas, pudiendo comparar la relación existente entre las variables; entre más cercano sea el valor a 1 significa que hay una fuerte relación entre las variables estudiadas, ya sea que una determina a otra o que son factor determinante en el aporte a la solución.

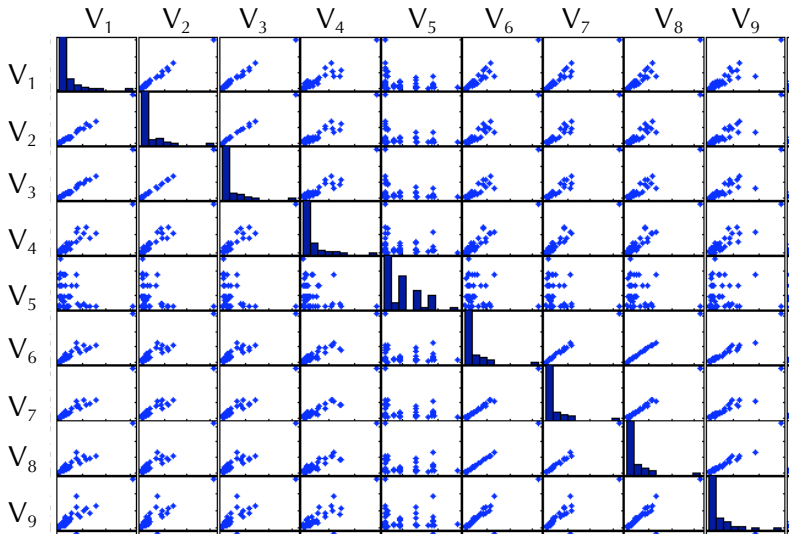

Figura 40. Diagrama de dispersión e histograma de las variables.

Para hacer más evidente la matriz de correspondencia de variables se generó un diagrama de dispersión e histograma, si observa la figura anterior, le será fácil determinar la relación existente entre las variables, ya que, entre más dispersos se encuentran los datos menor es relación existe entre las variables.

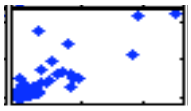

Figura 41. Dispersión en los datos; poca relación entre variables.

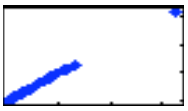

Figura 42. Relación estrecha entre variables, varianza cercana a 1.

7. Hasta este punto del Análisis de Componentes Principales podemos decir que ya se ha generado conocimiento, el cual consistió en la fabricación de *clusters*, agrupamientos de los cursos de cómputo de acuerdo a su efectividad.

   La representación del conocimiento se hizo a través de un dendograma, un tipo de gráfico que muestra los clusters encontrados en los datos a partir de la Distancia Euclidiana entre los mismos.

   Para hacer más evidente la forma en que se generó el dendograma, se obtendrán los componentes principales que reflejan la solución a la problemática analizada.

Un componente principal muestra en términos de varianza el porcentaje de explicación de un fenómeno. Es muy recurrente que para el Análisis de Componentes Principales se genere un gráfico utilizando los dos primeros componentes principales.

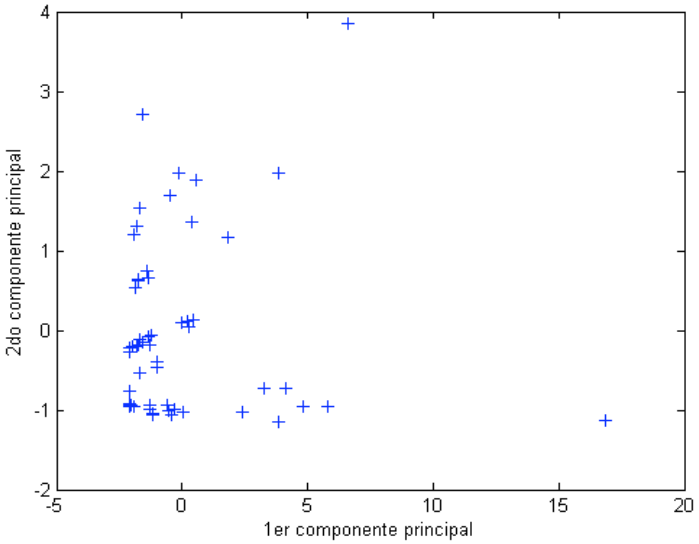

Figura 43. 1er y 2da componente principal.

El gráfico de la 1era y 2da componente principal muestra con cruces azules cada uno de los cursos que imparte la DCD de la DGSCA, UNAM; en este gráfico ya se ha reducido el universo de 12 a 1 sola variable, lo que permitió ligar los datos a través de la Distancia Euclideana.

Como se observa en la figura siguiente, existen 5 grupos (*clusters*) de grupos con una efectividad similar, mismo resultado obtenido que con el dendograma, a diferencia de que este último muestra grupos más estrechos y existe la posibilidad de ampliar la visión realizando agrupaciones en niveles superiores.

El primer y segundo componente generado deberán aportar un nivel de confianza suficiente para la explicación del fenómeno en cuestión, es decir deberán aportar más del 90% de de la solución del problema, de lo contrario, será necesaria la utilización de un tercer componente principal que aporte una mayor veracidad al fenómeno estudiado.

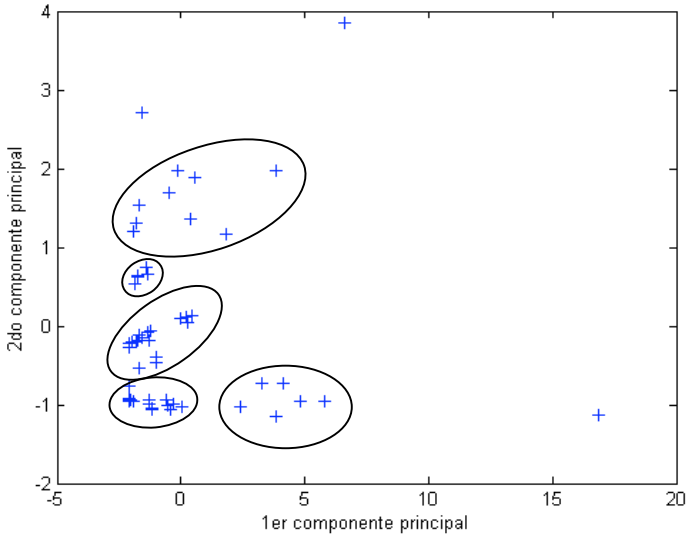

Figura 44. Reflejo de los clusters encontrados en los cursos de cómputo a partir de su efectividad.

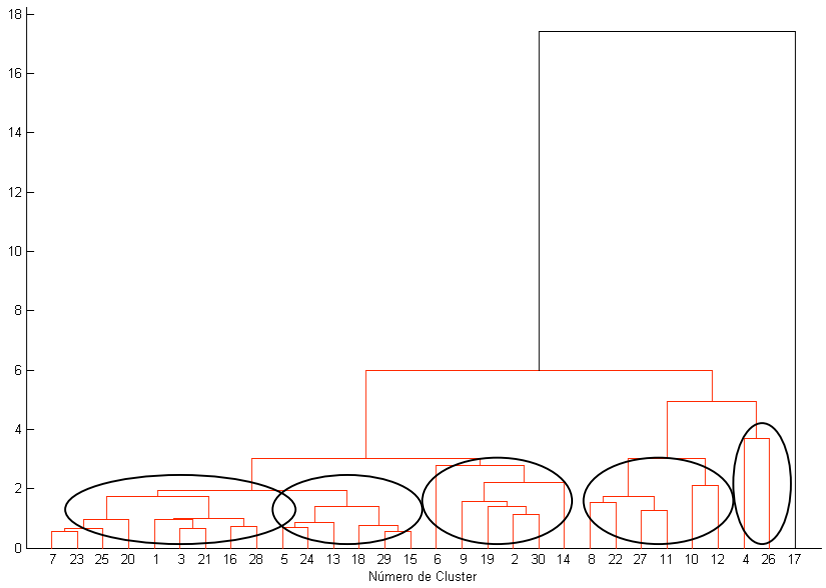

Figura 45. Cinco clusters principales y un caso especial.

8. Con el objeto de comprobar el nivel de confianza del estudio realizado, se procede a generar un gráfico en dónde se aprecie la aportación de cada componente principal en términos de la varianza de los datos, para que de esta forma se dictamine si el estudio cumple con las características suficientes para su aprobación.

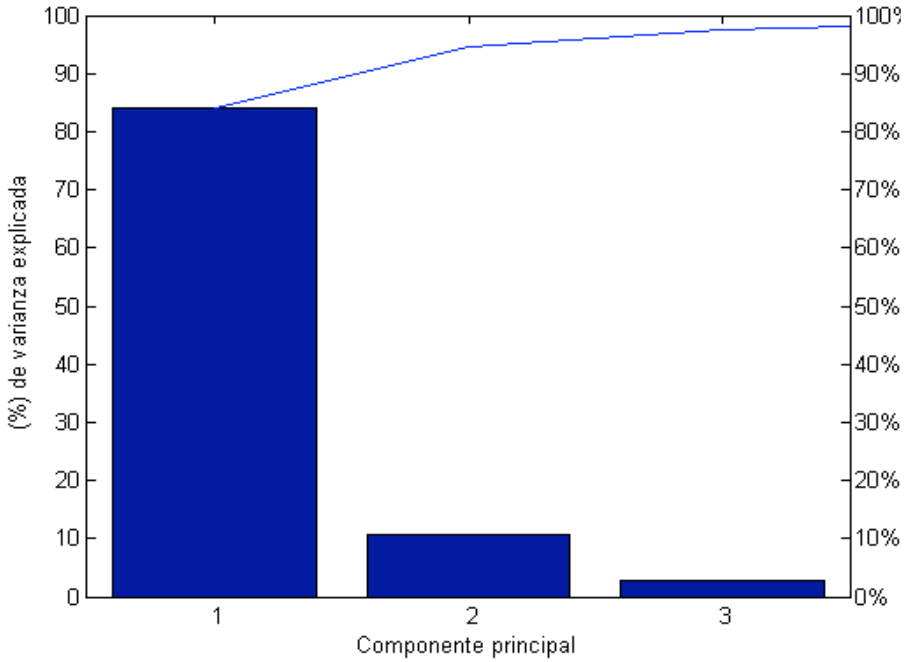

Figura 46. Aporte de solución por componente principal generado.

Al visualizar el gráfico de aporte de solución por componente principal, podemos gratamente apreciar que con la utilización de dos componentes principales la explicación del fenómeno estudiado se presenta en más del 90%, lo que hace al proceso de Minería de Datos válido y con un gran nivel de aceptación.

De utilizar tres componentes principales para dar explicación al fenómeno de la efectividad de los cursos de cómputo, se estaría alcanzando un nivel de confianza muy cercano al 100%, lo que da mayor aceptación al resultado obtenido.

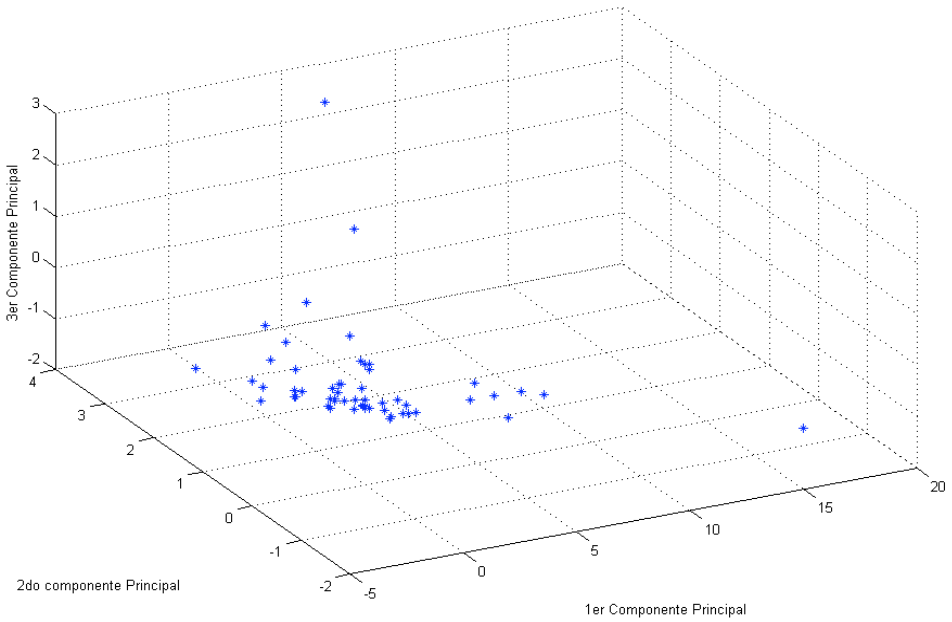

Figura 47. 1era, 2da y 3ra componente principal.

La figura anterior muestra el resultado que da explicación al fenómeno de estudio utilizando tres componentes principales en donde se encontraron los siguientes grupos (*clusters*).

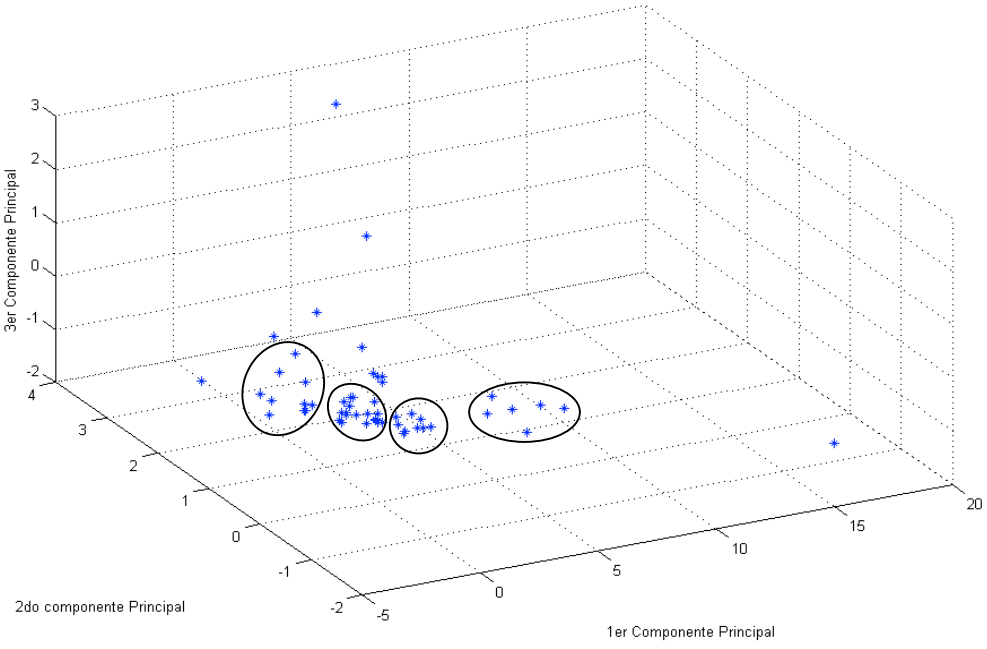

Figura 48. Clusters contenidos en la 1era, 2da y 3ra componente principal.

9. Como último punto se generará un índice que permita ordenar los cursos de cómputo de acuerdo a su efectividad, los clusters generados se utilizarán para realizar esta clasificación de cursos, ya que si recordamos el dendograma, los cursos con mayor efectividad se encuentran a la extrema derecha del gráfico, mientras que los de menor efectividad a la izquierda del gráfico.

   Lo que se realizará en esta tarea será un ordenamiento de los cursos de cómputo de acuerdo a la efectividad que representan, indicando el *cluster* o grupo al que pertenecen.

   La información que resultante de la generación del índice, representa el conocimiento obtenido luego de la implementación del Análisis de Componentes Principales, por lo cual resulta imposible su publicación ya que violaría el acuerdo de confidencialidad entre la DCD de la DGSCA, UNAM y el equipo de especialistas en Minería de Datos que realizó esta investigación.

   A manera de ejemplo se muestra la siguiente tabla de datos que reflejaría el conocimiento obtenido.

	Curso	Grupo (Cluster)
1	Minería de Datos	17
2	Análisis y Diseño de Sistemas	26
3	Word	26
4	Java	4
5	SQL Server	12
.		.
.		.
.		.
53	Internet	23
54	Crystal Reports	7
55	Power Point	7

Figura 49. Resultado final del análisis de componentes principales.

5.5.2. Evaluación de resultados por los dueños del negocio

Una vez realizada la implementación y evaluación del modelo de Minería de Datos, los resultados obtenidos del Análisis de Componentes Principales fueron presentados ante la L.I. Juana Figueroa Reséndiz titular de la Coordinación de Producción Académica quien realizó las siguientes observaciones:

- Los resultados presentados ante esta Coordinación, representan conocimiento nuevo y útil, en primera instancia debido a que no se contaba con los elementos necesarios para realizar una clasificación de los cursos que se imparten en la dependencia; en segundo término, conocer la efectividad que tienen los cursos de cómputo permitirá ejercer las acciones necesarias para realizar una adecuada calendarización de los mismos, con el objeto brindar el público en general la mejor oferta educativa en actualización de cómputo y tecnologías de la información.

- La generación de grupos de cursos de acuerdo a la efectividad resulta de gran utilidad en esta institución, ya que de esa forma se podrán ejercer acciones que permitan subsanar las deficiencias en los cursos donde se encontró una baja efectividad y con ello cumplir con el objetivo de la institución, que es la impartición de cursos de calidad.
- El listado de cursos que se obtuvo como resultado del proceso de Minería de Datos permitirá realizar una adecuada calendarización de los cursos de cómputo, ya que los cursos que tienen una baja efectividad podrán salir del calendario o bien, mejorar su estructura a fin de que alcancen un nivel de eficiencia óptimo, mientras que los cursos con alta efectividad podrán ser promocionados más ampliamente a fin de elevar sus niveles de eficiencia.
- Por el momento me es imposible ahondar en los beneficios que aporta el conocimiento adquirido del proceso de Minería de Datos ya que prefiero esperar a que el conocimiento se ponga en práctica y así evaluar los resultados de una manera adecuada. Cabe mencionar que al día de hoy ya se cuenta con una programación de cursos para los próximos 4 meses, por lo que la implementación del conocimiento se llevará a cabo hasta el mes de Noviembre de 2007.

5.5.3. Reseña del proceso

El estudio de Minería de Datos realizado para la Dirección de Cómputo para la Docencia de la DGSCA, UNAM, consistió en la realización de un análisis de *clusters*, a través de la implementación del Análisis de Componentes Principales.

Para llevar acabo dicho estudio, se tomó como base la metodología CRISP-DM la cual implementa un ciclo de vida para Minería de Datos, en donde se atraviesa un conjunto de 6 fases:

- Comprensión del negocio.

  En esta fase el objetivo principal consistió en lograr la comprensión de los objetivos y requerimientos del proyecto desde la perspectiva particular del negocio, con el fin de traducir el conocimiento adquirido en una definición del problema de Minería de Datos y en el diseño de un plan preliminar. Se realizaron diversas entrevistas con el personal de la institución a fin de conocer las expectativas del proceso de Minería de Datos a ejecutar.

- Comprensión de los datos.

  En esta fase las actividades realizadas consistieron en la obtención de los datos, su familiarización y evaluación de la naturaleza de los mismos con el fin de distinguir las variables a utilizar, de identificar las agrupaciones posibles y sobre todo de plantear una hipótesis inicial.

- Preparación de los datos.

  La fase de preparación de datos consistió en la realización de actividades que permitieron la construcción del conjunto de datos a utilizar durante el proyecto; las tareas de ésta fase incluyeron la selección de datos a utilizar, generación de variables, agrupación de datos y limpieza de los mismos.

  Es preciso recalcar el hecho de que esta fase en particular tuvo una duración del 70% del total del tiempo invertido en el proyecto, conformar el conjunto de datos a utilizar resulta un trabajo arduo ya que se deben seleccionar los elementos que mejor describan la situación del fenómenos a estudiar, así mismo, la limpieza y homogenización de valores resulta complicado cuando se cuenta con grandes volúmenes de datos.

- Modelado

  En esta fase se seleccionó de varias técnicas la más óptima para cumplir con el objetivo de Minería de Datos, así pues, se eligió la realización de un análisis de *clusters* a través del Análisis de Componentes Principales, la implementación del mismo se realizó en el software matemático y estadístico MatLab, también se definieron los criterios de evaluación del modelo para su correcto funcionamiento.

- Evaluación

  Durante esta fase del proyecto, se llevó a cabo la evaluación del modelo construido con el conjunto de datos generado en fases anteriores, la evaluación del modelo consistió en un recuento paso a paso del proceso realizado para el Análisis de Componentes Principales, obteniendo como resultado final el conocimiento esperado por los dueños del negocio.

- Despliegue

  La fase de despliegue consistió en la elaboración de un plan que permite evaluar el conocimiento adquirido en la fase anterior, a su vez se prevé un plan de mantenimiento para el modelo generado en MatLab.

5.6. Despliegue

5.6.1. Plan de despliegue

Como fase final del proceso de Minería de Datos nos encontramos con un plan de despliegue, que en sí engloba un conjunto de recomendaciones y acciones a seguir para la implementación del conocimiento adquirido en el Modelo de Negocio que sigue la Dirección de Cómputo para la Docencia de la DGSCA, UNAM.

Así pues las observaciones y recomendaciones que hace el equipo de trabajo encargado del proceso de Minería de Datos a la institución, son las siguientes:

- La información analizada durante el proceso de Minería de Datos, resultó sustancial para el Análisis de Componentes Principales, bajo este marco de referencia es como se generó un conjunto de grupos de cursos de cómputo, en donde el elemento que permitió realizar dicho agrupamiento fue la efectividad de un curso de cómputo en términos económicos y académicos.
- Se encontraron en principio 30 grupos de cursos diferentes, logrando una minimización de los mismos obteniendo como resultado final 5 grupos de cursos yendo estos de mayor a menor efectividad.
- Los cursos que alcanzaron una menor efectividad requieren de un seguimiento especial por parte de las autoridades de la institución, ya que, son cursos que no satisfacen al público en general y que económicamente resultan poco rentables para la institución, así pues estos cursos deberán mejorar su estructura académica, es decir se deberá evaluar el contenido de los mismos a fin de determinar si es que cumplen con las expectativas de los clientes (alumnos).

  Una vez que se ha evaluado y mejorado la estructura del contenido de los cursos, es como se podrá realizar una adecuada calendarización de los mismos, teniendo en cuenta que al modificar la estructura del curso, cambiará el costo de inscripción al mismo, el objetivo principal es agregar valor académico a los cursos de baja efectividad para poder así replantear el precio de

salida al mercado y con ello, lograr un estimulo en el mercado que permita una mayor captación de clientes (alumnos).

- Entendiendo que los elementos tecnológicos se encuentran en constante actualización es como se recomienda que los cursos que obtuvieron una efectividad alta, no sean descuidados, estar a la vanguardia en conocimiento científico y tecnológico es un valor agregado para la institución, por lo que para conservar la efectividad de los cursos, se sugiere que se monitoree constantemente el contenido de los mismos.
- La calendarización que se hace de los cursos donde se encontró una gran efectividad es adecuada, se sugiere se amplíen los medios publicitarios para este tipo de cursos, con el objeto de logran una mayor captación de público interesado en los mismos.

5.6.2. Plan de monitoreo y mantenimiento

El plan de monitoreo y mantenimiento está enfocado a emitir un conjunto de recomendaciones que permitan replicar el proceso de Minería de Datos en un futuro.

1. La preparación de los datos es un proceso determinante para el éxito del estudio de Minería de Datos realizado; el equipo de trabajo conformó un conjunto de vistas actualizables dentro de la base de datos de SAEC[20] a manera de un pequeño Dataware House, el cual tiene como objetivo presentar los datos que serán enviados al modelo implementado en MatLab. Dentro de este conjunto de vistas, existe una en especial que deberá ser monitoreada constantemente a fin de realizar un tratamiento sobre valores nulos y datos que requieren de actualización, la vista en cuestión esta nombrada como *"V_ValoresNulos"*.
2. Una vez que se obtuvo el DataSet a utilizar para llevar a cabo el Análisis de Componentes Principales, se deberá validar que los valores contenidos para las variables de tipo texto tengan una longitud uniforme, ya que es un requisito del modelo implementado.
3. La implementación del Análisis de Componentes Principales se realizó en el software matemático y estadístico MatLab, la librería generada se encuentra a espera de un DataSet de dimensiones *n x n*, lo que lo hace flexible ante cualquier situación que se presente, no se requerirá más que enviar el DataSet para la generación de resultados.
4. La interpretación de los gráficos y demás salidas que proporciona el Análisis de Componentes Principales deberá realizarse por un experto en estadística descriptiva y multivariante, se dará un pequeño entrenamiento de esta situación a las personas que la DCD de la DGSCA, UNAM crea conveniente.

[20] Sistema de Administración de Educación Continua.

5.6.3. Elaboración del reporte final

México D.F., a 6 de julio de 2007

Asunto: Reporte final sobre la culminación del proyecto de Minería de Datos CComp01.

M. en C. Jesús Díaz Barriga Arceo.
Subdirector de Planeación Académica
DCD, DGSCA, UNAM

P r e s e n t e

Por medio del presente documento permítame hacer de su conocimiento que el proyecto de Minería de Datos realizado sobre la base de datos del Sistema de Administración de Educación Continua SAEC ha llegado a su fin en términos satisfactorios.

El estudio realizado consistió en la evaluación de los cursos de cómputo impartidos por la dependencia para el periodo comprendido entre el 1 de enero de 2006 y 15 de junio de 2007, a fin de determinar la efectividad de los mismos en términos académicos y económicos.

El estudio de Minería de Datos realizado permitió identificar de forma clara cuál es posicionamiento de los cursos que son impartidos en los centros de extensión de la DGSCA, lo cual permitirá realizar una óptima calendarización de cursos de cómputo en un futuro.

El objetivo planteado por la institución consiste en allegarse de un mayor número de recursos y disminuir el número de cursos programados sin impartir, a través de una adecuada calendarización de cursos de cómputo; dicha calendarización se pondrá en marcha a partir del mes de Noviembre del año en curso, los resultados observables estarán a la vista al término del periodo de inscripciones.

Dentro del estudio de Minería de Datos se obtuvo como resultado un listado de cursos de cómputo cuyo orden es la efectividad que representan para la institución, así mismo se generaron agrupamientos de cursos que permiten analizar la situación de los mismos por bloque, y a su vez, es lo que permitirá tomar las acciones pertinentes para mejorar la efectividad de los mismos.

El conocimiento generado se ha puesto a su disposición en la documentación del proyecto previamente proporcionada. Así mismo permítame hacer de su conocimiento que se ha implementado un sistema semiautomático el cual le permitirá a la Institución replicar el estudio de Minería de Datos en un futuro. Se hizo entrega de una pequeña aplicación desarrollada en el software matemático y estadístico MatLab, la cual se encuentra alojada en uno de los servidores de la Institución, el funcionamiento de la misma se explica en la documentación del proyecto.

La validez de los estudios realizados cuenta con un nivel de confianza mayor al 90%, el cual se había determinado como objetivo para el proyecto de Minería de Datos, las pruebas realizadas fueron sustanciales y me permiten afirmar que no existen errores en los resultados proporcionados.

También es importante mencionar el hecho de que la base de datos con la cual se trabajó en el proyecto es susceptible de una explotación mayor, se encontraron los elementos suficientes para llevar a cabo nuevos estudios de Minería de Datos que aporten un mayor conocimiento a la Institución que usted representa.

El equipo de trabajo a cargo del proyecto de Minería de Datos trabajó siempre bajo los lineamientos establecidos en el acuerdo de confidencialidad firmado al inicio del proyecto, de tal modo que en ningún momento se comprometió siquiera la información referente al Modelo de Negocio que sigue la DCD de la DGSCA, UNAM.

Sin más por el momento me despido, quedando de usted en la mejor disposición.

Atte. Carlos Tomás Reyes García.
Coordinación de Producción Académica

Capítulo 6

Conclusiones

A lo largo del presente documento se ha discutido sobre el valor que tiene la información para las organizaciones; el hombre desde el principio de los tiempos se ha preocupado por transmitir su conocimiento a generaciones futuras con el simple objeto de lograr una evolución de la especie.

Hoy en día, entendiendo que vivimos rodeados por un mundo globalizado, las organizaciones tienden a generar grandes volúmenes de información, misma que es captada día a día por sistemas computacionales, hecho que refleja la historia de las organizaciones, historia que permitirá su evolución.

Una organización, al igual que el hombre, requiere conocer de donde viene, cual es su historia, para así, dirigirse al lugar que anhelan llegar; lamentablemente para las organizaciones esta tarea no resulta sencilla, ya que como hemos mencionado, una organización genera grandes volúmenes de información a través del tiempo y su apreciación sobre ésta, comienza a ser cada vez más abstracta; recordar el detalle de todas las operaciones efectuadas no es una tarea sencilla y es precisamente ahí donde la Minería de Datos interviene, ya que el proceso generado por esta área de conocimiento permite explotar grandes volúmenes de información para generar conocimiento entendible y útil que una organización deja de apreciar por la gran cantidad de datos que maneja.

Este documento de tesis cumple con su propósito al ilustrar de manera clara que cuando se cuenta con un gran volumen de información es posible su explotación a través de técnicas estadísticas, matemáticas y computacionales, con el objeto de hallar patrones ocultos en los datos que permitan la generación de conocimiento, en donde a su vez, éste dé pauta a las organizaciones para guiar su camino hacia una evolución promisoria.

Explotar grandes volúmenes de información es una tarea difícil, generar conocimiento lo es aún más, ya que existe una gran diversidad de técnicas para llevar a cabo un proceso de Minería de Datos, en donde aumenta el grado de complejidad sabiendo que para lograr esto se requiere de la integración de áreas de conocimiento como la matemática, estadística e informática que por sí mismas representan grandes áreas de conocimiento.

Una vez finalizada la investigación y llevado a cabo un caso de aplicación práctico sobre Minería de Datos es como podemos concluir que:

1. El hombre se ha preocupado desde el principio de los tiempos por conservar su historia y transmitir su conocimiento a generaciones futuras con el objeto de lograr una evolución en la especie; hecho que no resulta consciente para todos los hombres.
2. Las organizaciones al igual que el hombre cuentan con una historia, misma que les permite tomar las decisiones que conlleven a su evolución, sin embargo cuando la historia de una organización es amplia, los volúmenes de información impiden a las organizaciones la visión sobre el detalle de las operaciones, dejando de lado información valiosa que a futuro puede traducirse en conocimiento.

3. La Minería de Datos es un proceso a través del cual grandes volúmenes de información son explotados a través de técnicas estadísticas, matemáticas y computacionales, con el objeto de hallar patrones ocultos en los datos y así generar conocimiento entendible y útil para las organizaciones.
4. Para la realización de un estudio de Minería de Datos como mínimo hay que atravesar por las siguientes fases:
    a. Comprensión de la situación actual.
    b. Definición de objetivos.
    c. Elección sobre el modelo de Minería de Datos a utilizar.
    d. Preprocesamiento de datos.
    e. Implementación del modelo.
    f. Análisis de resultados.
    g. Presentación del conocimiento generado.
5. Las técnicas de Minería de Datos son muy diversas, saber cuál es la adecuada para cada caso de estudio dependerá en primera instancia de los objetivos que del proyecto se señalen y posteriormente de los elementos con que se cuenten para la realización del estudio.
6. La fase de preprocesamiento de datos es la que más tiempo consume del proyecto, hasta un 70% del mismo, ya que, es el factor determinante en la mayoría de los casos para que el estudio de Minería de Datos se lleve exitosamente.

La realización de un caso de estudio práctico permitió asentar cada una de las aseveraciones anteriores, se estudió el fenómeno de impartición de cursos de cómputo, en donde a través de un análisis de componentes principales se generó un indicador que explica en términos de efectividad el posicionamiento de los cursos de cómputo impartidos por la DCD de la DGSCA, UNAM.

Así mismo, la implementación de un caso de estudio práctico de Minería de Datos en la DCD de la DGSCA, UNAM, permite concluir que:

1. La Minería de Datos si puede ser aplicada como una herramienta para la toma de decisiones en el proceso de calendarización de cursos de cómputo. La Coordinadora de Producción Académica de la DCD, de la DGSCA UNAM menciona que:
    a. La generación de grupos de cursos de acuerdo a la efectividad resulta de gran utilidad en la institución, ya que de esa forma se podrán ejercer acciones que permitan subsanar las deficiencias en los cursos donde se encontró una baja efectividad y con ello cumplir con el objetivo de la institución, que es la impartición de cursos de calidad.
    b. El listado de cursos que se obtuvo como resultado del proceso de Minería de Datos permitirá realizar una adecuada calendarización de los cursos de cómputo, ya que los cursos que tienen una baja efectividad podrán salir del calendario o bien, mejorar su estructura a fin de que alcancen un nivel de eficiencia óptimo, mientras que los cursos con alta efectividad podrán ser promocionados más ampliamente a fin de elevar sus niveles de eficiencia.
2. El Análisis de Componentes Principales realizado a la Dirección de Cómputo para la Docencia permite sentar las bases para realizar análisis ulteriores de Minería de Datos; el análisis de Minería de Datos inmediatamente factible consiste en la generación de *clusters* de alumnos con el objeto de realizar predicciones sobre los grupos de alumnos susceptibles a inscribirse a cierto tipo de cursos.

Bibliografía

- ***Catherine M., Ricardo***. *2004*. Databases Illuminated*, Mississagua, Canadá: Jones and Barlett Publishers*.
- ***Han, Jiawei*** *y* ***Micheline Kamber***. *2001*. Data Mining Concepts and Techniques, *M. Kaufmann (ed.), San Francisco, CA: Academic Press*.
- ***Joyanes Aguilar, Luis y Zahonero Martínez Ignacio.*** *1998*. Estructura de Datos: Algoritmos, abstracción y objetos. *Madrid, España. Mc Graw Hill*.
- ***Berry, Michel*** *y* ***Linoff, Gordon***. *2004*. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. *Indianapolis, Indiana: Wiley Publishing, Inc*.
- ***Witten, Ian H.*** *y* ***Eibe, Frank.*** *2005*. Data Mining Practical Machine Learning Tools and Techniques. *San Francisco, CA: Elsevier*.
- ***Larose, Daniel T.*** *2005*. Discovering Knowledge in Data, An Introduction to Data Mining. *New Jersey: Wiley Publishing, Inc*.
- ***Kimball, Ralph*** *y* ***Ross, Margy***. *2002*. The Data WareHouse Toolkit. *New York, EUA: Wiley Publishing, Inc*.
- ***Chapman, Pete – Clinton, Julian – Kerber, Randy – Khabaza, Thomas – Reinartz, Thomas – Shearer, Colin*** *y* ***Wirth, Rüdiger.*** *2000*. CRISP – DM 1.0 Step-by-step data mining guide. *EUA: CRISP – DM consortium*.
- ***González, Pilar – Díaz, Amelia – Torres, Enrique – Garnica, Elsy.*** *1994. "Una aplicación del análisis de componentes principales en el área educativa", en* Revista Economía [Universidad de los Andes, Instituto de Investigaciones Económicas y Sociales]*, 1994:Núm. 9. Pág. 55 – 72*.
- ***Aluja Banet, Tomas.*** *1999*. Aprender de los datos: el análisis de componentes principales: una aproximación desde el data mining. *Barcelona. EUB*.
- ***Van Der Lans, Rick F***. *2006*. Introduction to SQL: Mastering the Relational Data Base Language. *Estados Unidos de Norte America: Addison Wesley Professional*.
- ***Taylor, Allen G.*** *2006*. SQL for Dummies. *Indianapolis, EUA. Wiley Publishing, Inc*.
- ***Donahoo, Michel J. y Speegle, Gregory D.*** *2005*. SQL: Practical Guide for Developers. San Francisco*, UEA. Ed. Morgan Kauffman Ed., Elseiver Inc*.
- ***[Web 01.] Wikimedia Foundation, Inc***. Papel, *en http://es.wikipedia.org/wiki/Papel. Visitada el 4 de abril de 2007*.
- ***[Web 02.] Wikimedia Foundation, Inc***. Cáncer, *en http://es.wikipedia.org/wiki/C%C3%A1ncer. Visitada el 4 de abril de 2007*.
- ***[Web 03.] Wikimedia Foundation, Inc***. Almacén de datos, *en http://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos. Visitada el 1 de julio de 2007*.
- ***[Web 04.] Pisanty Baruch, Alejandro***. Quienes somos*, en http://www.dgsca.unam.mx/somos.html. Visitada el 15 de febrero de 2007*.