

Universidad Autónoma de Baja California

Facultad de Contaduría y Administración

Facultad de Ciencias Químicas e Ingeniería

Maestría en Tecnologías de la Información y la Comunicación



**Implementación de minería de datos para apoyo en la toma de decisiones en
la agencia automotriz Renault.**

Para obtener el grado de

Maestro en Tecnologías de la Información y la Comunicación

Presenta:

Bernal Osorio Yajaira

Bajo la dirección de:

Dra. María del Consuelo Salgado Soto

Agradecimientos

Gracias a esas personas importantes en mi vida, que siempre estuvieron listas para brindarme toda su ayuda, ahora me toca regresar un poquito de todo lo inmenso que me han otorgado. Con todo mi cariño esta tesis se las dedico a ustedes:

A mi madre, que con todo su cariño y amor hizo todo en la vida para que yo pudiera lograr mis sueños, por motivarme y darme la mano cuando sentía que el camino se terminaba, por siempre mi amor y agradecimiento hacia ella.

A mi padre, por su apoyo y cariño, que aunque no esté físicamente ve por mis sueños.

Madrina Marie Laure, por su sabiduría, siempre brindarme su apoyo, por confiar, creer en mí y haber estado siempre en cada etapa y trayecto de mi vida.

Sobre todo a Dios por haberme permitido vivir hasta este día, haberme guiado a lo largo de mi vida, se mi apoyo, mi luz y camino. Darme fuerzas para continuar adelante en aquellos momentos difíciles.

A ti mi compañero por tu paciencia y comprensión, por sacrificar nuestros tiempos para que pudiéramos cumplir nuestros compromisos. Por tu bondad y sacrificio, gracias por estar a mi lado.

A mis maestros y directora de tesis Dra. Consuelo Salgado, que en este andar por la vida, influyeron con sus lecciones y experiencias en formarme como una persona de bien y preparada para los retos que pone la vida, a todos y cada uno de ellos les dedico cada una de estas páginas de mi tesis.

A mis amigos, por todos los momentos que pasamos juntos por los nuevos y viejos amigos que siempre están apoyando, brindando su tiempo y perseverancia. Por la confianza y cariño que me continúan brindando.

INDICE

INDICE DE FIGURAS.....	6
INDICE DE TABLAS	7
CAPÍTULO I. INTRODUCCIÓN	8
1.1 INTRODUCCIÓN.....	9
1.2 ANTECEDENTES.....	10
1.3 PLANTEAMIENTO DEL PROBLEMA	14
1.4 JUSTIFICACIÓN	15
1.5 OBJETIVOS.....	16
<i>Objetivo General.</i>	<i>16</i>
<i>Objetivos Específicos.</i>	<i>16</i>
1.6 ALCANCES	17
1.7 LIMITACIONES.....	17
CAPÍTULO II. MARCO TEÓRICO	18
2.1 MINERÍA DE DATOS	19
2.1.1 <i>Definición.....</i>	<i>19</i>
2.1.2 <i>Descripción del proceso de minería.....</i>	<i>21</i>
2.1.3 <i>Definición de cubo de datos.....</i>	<i>22</i>
2.1.4 <i>Clasificación</i>	<i>22</i>
<i>Regresión.....</i>	<i>23</i>
2.1.5 <i>Áreas relacionadas con la minería de datos</i>	<i>23</i>
2.1.6 <i>Origen de la minería de datos.....</i>	<i>24</i>
2.1.8 <i>DMQL (Lenguaje de minería de datos de consulta).....</i>	<i>29</i>
2.1.9 <i>Minería de Datos para la toma de decisiones en las actividades de negocio....</i>	<i>30</i>
2.2 METODOLOGÍAS DE MINERÍA DE DATOS	32
2.2.1 <i>CRISP-DM (Cross Industry Standard Process for Data Mining).....</i>	<i>32</i>
2.2.2 <i>SEMMA (Sample, Explore, Modify, Model and Assess).....</i>	<i>36</i>
2.2.3 <i>Diferencias entre SEMMA - CRISP-DM.....</i>	<i>40</i>
2.3 HERRAMIENTA DE MINERÍA DE DATOS.....	43
2.3.1 <i>Definición.....</i>	<i>43</i>
2.3.2 <i>Características de RapidMiner.....</i>	<i>43</i>
2.4 TOMA DE DECISIONES	45

2.4.1 Definición.....	45
2.4.2 La toma de decisiones y la organización del trabajo en la empresa.....	46
2.4.3 Etapas en el proceso de toma de decisiones.....	49
2.4.4 Sistema de aprendizaje	52
2.5 AGENCIA AUTOMOTRIZ.....	53
2.5.1 Definición.....	53
2.5.2 Guía de Operación.	53
2.5.3 Preparación pre-entrega de una unidad nueva.....	54
2.5.4 Servicio de las millas u otra “revisión”	54
2.5.5 Visita de garantía.....	55
2.5.6 Primera visita post garantía.	55
CAPITULO III. METODOLOGÍA.....	56
3.1 FASES DE LA METOLOGIA CRISP-DM	58
3.1.1 Comprensión del negocio	58
3.1.2 Fase de comprensión de los datos	59
3.1.3 Fase de preparación de los datos.....	60
3.1.4 Fase de modelado.....	61
3.1.5 Fase de evaluación	61
3.1.6 Fase de implementación.....	62
CAPÍTULO IV. DESARROLLO	64
4.1 COMPRENSION DEL NEGOCIO.	65
4.2 COMPRENSIÓN Y ANÁLISIS DE DATOS	68
4.3 PREPARACIÓN DE DATOS	75
4.4 MODELADO	82
CAPÍTULO V. RESULTADOS.....	83
CAPITULO VI. CONCLUSIONES.....	83
BIBLIOGRAFÍA.....	83
ANEXOS	83

INDICE DE FIGURAS

CAPÍTULO I. INTRODUCCION

Figura 1. 1 Organigrama general de agencia automotriz Renault.....	11
--	----

CAPÍTULO II. MARCO TEORICO

Figura 2.1. Proceso de la minería de datos (Gilberto Lorenzo Martínez Luna, Adolfo Guzmán Arenas, 1994)	19
Figura 2.2. Variables de las bases de datos de sistemas operacionales (Gilberto Lorenzo Martínez Luna, Adolfo Guzmán Arenas, 1994)	22
Figura 2.3. Dimensiones o ejes que forman los cubos y por los cuales los mineros se desplazarán (Gilberto Lorenzo Martínez Luna, Adolfo Guzmán Arenas, 1994).....	22
Figura 2.4 Niveles de Abstracción.....	33
Figura 2.5 Fases de metodología SEMMA.....	37
Figura 2.6 Dinámica de la Metodología SEMMA. Extraído de (Vendrell, 2011)	37
Figura 2.7 Gráfica comparativa (Pérez López, C. y Santin González, D., 2007).....	40
Figura 2.8 Toma de decisiones en el entorno actual.....	48
Figura 2.9 Sistema de aprendizaje.....	52

CAPÍTULO II. METODOLOGÍA

Figura 3.1 Ciclo de vida de la metodología CRIPS-DM (Fuente Propia).....	58
Figura 3.2 Fase 1 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	57
Figura 3.3 Fase 2 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	58
Figura 3.4 Fase 3 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	59
Figura 3.5 Fase 4 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	60
Figura 3.6 Fase 5 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	61
Figura 3.7 Fase 6 del Ciclo de vida de la metodología CRIPS-DM (Fuente Propia)	62

INDICE DE TABLAS

CAPÍTULO II. MARCO TEÓRICO

Tabla 2. 1 Comparativo de metodologías.....	41
---	----

CAPÍTULO IV. DESARROLLO

Tabla 4.1: Muestra atributos a utilizar	67
Tabla 4.2: Muestra campo con su tipo de dato.....	69
Tabla 4.3: Datos seleccionados para el análisis.....	73
Tabla 4.4: Tabla empresa con sus atributos.....	77
Tabla 4.5: Tabla detalle venta con sus atributos	77
Tabla 4.6: Tabla producto con sus atributos.....	77
Tabla 4.7: Tabla atributosprod con sus atributos.....	78
Tabla 4.8: Tabla color con sus atributos.....	78
Tabla 4.9: Tabla modelo con sus atributos.....	78
Tabla 4.10: Tabla almacén con sus atributos	78
Tabla 4.11: Tabla vendedor con sus atributos.....	78
Tabla 4.12: Tabla cliente con sus atributos	79
Tabla 4.13: Tabla venta con sus atributos.....	79
Tabla 4.14: Clúster por versión	86
Tabla 4.15: Representación periodo 2012 Tijuana, por versión	90
Tabla 4.16: Representación periodo 2013 ventas por versión de unidad.....	92

INDICE DE GRÁFICAS

CAPÍTULO IV. DESARROLLO

Gráfica 1: Desconocimiento de unidades más vendidas.....	69
Gráfica 2: Tipo unidad según modelo.....	70
Gráfica 3: Cantidad de unidades según modelo.....	70
Gráfica 4: Cantidad de unidades según modelo.....	71

CAPÍTULO I.

INTRODUCCIÓN

1.1 Introducción

La minería de datos es una tecnología con un gran potencial para ayudar a los negocios a enfocarse en la información más importante de las bases de datos o almacenes de datos, estas herramientas predicen comportamientos permitiendo a los gerentes y directivos crecer en tomar de decisiones y manejo de conocimiento. Una perspectiva automatizada de un análisis en minería va más allá de eventos pasados y puede responder a preguntas gerenciales de forma más rápida que antes al consumir más tiempo en responder.

En la actualidad, la velocidad en que la información se almacena es superior a la velocidad en que es analizada, es decir, que alrededor del mundo se ha estimado que el crecimiento de las bases de datos se va duplicando cada cierto tiempo, a diferencia de las técnicas de análisis de información. La minería de datos puede considerarse como una técnica para el análisis de información cuando se tiene la necesidad y el interés de explotar grandes volúmenes, y al desconocer la forma en que pueda transformarse en conocimiento que apoye en la toma de decisiones en especial a un nivel gerencial.

Este proyecto presenta estrategias que pueden favorecer a un departamento para incrementar sus ganancias y poder mantenerse en un ámbito competitivo, al identificar estrategias claves para satisfacer las necesidades en el mercado, para lograrlo se requiere una tecnología que permita tener información efectiva y segura en una base de datos.

Este trabajo de investigación se organiza en cinco capítulos; en el primer capítulo se encuentra el planteamiento del problema, los objetivos, la justificación, limitación y alcance del estudio. El segundo capítulo, contiene la teoría que fundamenta el proyecto.

El capítulo tercero, describe las fases que se desarrolla el proyecto. En el capítulo cuarto se presentan los resultados de la herramienta aplicada. Y por último, se presentan las conclusiones y recomendaciones.

1.2 Antecedentes

La minería de datos ha sido implementada en investigaciones ya sea para toma de decisiones, apoyo de estrategias, consulta de información al momento, por ser una herramienta que se adapta y sobre todo muestra resultados exactos, los cuales benefician al investigador y a la empresa involucrada, ya que no solo se enfoca en el ámbito tecnológico sino que puede implementarse en cualquier ramo.

Hay una variedad de metodologías que facilitan el proceso de su uso, como software instalable donde, a través de la comunicación del manejador de base de datos se crea una comunicación por medio de una serie de pasos donde el mismo software optará en desplegar la información deseada, por ejemplo, gráficas en las cuales se puede observar comparativos ya sea por tiempo, cantidad, entre otras.

La minería de datos o explotación de datos se puede definir como un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y comprensible, a partir de grandes conjuntos de datos con motivo de predecir de forma automatizada tendencias, comportamientos y/o descubrir modelos previamente desconocidos.

Desde un punto de vista empresarial, la minería de datos puede ser definida como un conjunto de áreas que tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten información hacia la toma de decisiones; la finalidad es reunir las ventajas de varias áreas como la estadística, inteligencia artificial y computación gráfica, donde las bases de datos son la materia prima. (Aguilar, 2011)

El realizar las consultas convencionales a grandes bases de datos no es suficiente para resolver problemas de negocios, sino que se hace necesario seguir una metodología ordenada para aplicar herramientas tecnológicas y técnicas disponibles para así obtener conocimiento y resultados que permitan a las compañías obtener un beneficio, esta trata de sacar información de detalle de un

almacén de datos y no se conforma sólo con la visualización de datos, sino que es necesario obtener resultados en cuanto a la relación que existe entre los mismos datos y como podrían repercutir en el negocio.

En la actualidad, los negocios como las agencias automotrices necesitan estar a la vanguardia para mantenerse en un ámbito competitivo, el poder identificar estrategias claves para satisfacer las necesidades del mercado basándose en un estudio comparativo, donde requieren una tecnología que permita tener información veraz y segura en una base de datos.

La investigación se llevó en la agencia automotriz Renault, ubicada en los estados de Baja California y Sonora que cuenta con cuatro sucursales, tres en Baja California, dispersas de la siguiente manera: una en Tijuana y dos en Mexicali, en Sonora se encuentra una agencia en el municipio de Hermosillo. La estructura de la organización se encuentra conformada por los departamentos funcionales de Sistemas, Servicios, Ventas y Administración como muestra la Figura 1.1. Cada uno cuenta con un área en específico y encargadas de realizar sus tareas, mantener comunicación entre los mismos departamentos para la toma de decisiones, que deben garantizar la correcta funcionalidad de los diferentes sistemas necesarios para la operatividad de la empresa además de contar con una estructura funcional en su giro.

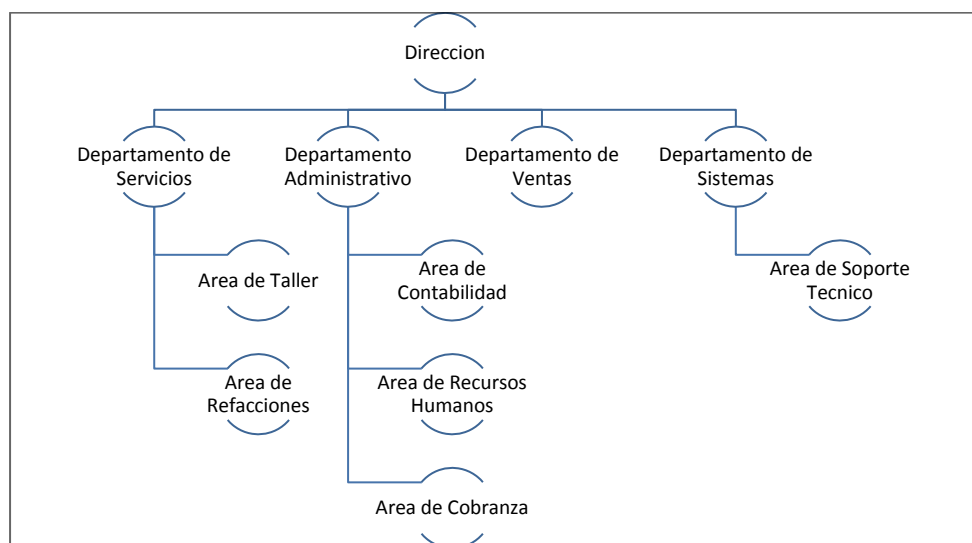


Figura 1. 1 Organigrama General de Agencia Automotriz Renault (Fuente Propia)

A inicios del 2012, la agencia automotriz Renault realizó un estudio donde observó que las ventas necesitaban mejorarse en comparación con periodos anteriores y la falta de herramientas que beneficiarán en la toma de sus decisiones, motivo para iniciar una investigación para obtener resultados estadísticos de los comparativos entre periodos y poder favorecer a la agencia.

La agencia decidió optar por una herramienta que beneficiará a la misma en la extracción de información de interés de la gran base de datos que manejaba, por tal motivo la agencia se dio a la tarea de realizar un análisis para poder verificar si en efecto la disminución de las ventas continuaban, iniciaron con la aplicación de una encuesta a los clientes para verificar la calidad del servicio que recibían, no solo se observó la reducción de venta de unidades sino también la disminución de venta de refacciones, se realizó un estudio con la información que arrojaba el sistema, se verificó que, en efecto la reducción de estas era inminente.

En los últimos meses del año del 2013, el departamento de ventas de la agencia automotriz Renault ha tenido un resultado negativo en sus ventas, por lo que se requiere optar por estrategias para encontrar una solución viable, la cual pueda ofrecer tener información de periodos anteriores para realizar comparativos y optar por una mejor estrategia de venta.

La utilización de una herramienta como la minería de datos proporcionará un panorama donde se podrá determinar un modelo o escenarios de patrones a través de datos históricos para situar tendencias y comportamientos, que se suscitan en la agencia, ya que se desea obtener un comparativo de las ventas reflejadas por períodos y vendedores en años anteriores para aplicar estrategias en el negocio favorables para una mejora continua. Se cuenta con una base de datos para el almacenamiento de la información, en la cual se procesan sus datos y se dispone de ellos en un momento oportuno, por tal motivo se realiza esta investigación para obtener un mejor manejo de la información con apoyo de una herramienta se pueda extraer la información de las bases de datos, basándose en técnicas de procesamiento de datos, tecnología y sistemas de información, así como realizar un análisis detallado con el objetivo de conocer patrones y

secuencias, en apoyos a la toma de decisiones con un menor grado de incertidumbre.

Uno de los estudios consultados fue implementado por Ing. Hugo Daniel Flores (2009), que se enfocó en la detección de patrones de daños y averías en la industria automotriz, utilizó la metodología CRIPS-DM para detectar patrones en la producción de daños en la cadena de distribución y esto permitió a la industria identificarlos y determinarlos como averías y su gravedad. El tener identificado los lugares donde se producen las averías y poder definir patrones de comportamiento con la utilización de la metodología antes mencionada y la minería de datos, le fue posible determinar patrones estadísticos aplicados en su agencia automotriz, una vez establecido que su información estaba normalizada fue posible su estudio donde logró la agrupación de datos para ser procesados por distintas herramientas de software y obtener resultados para los interesados. En la minería de datos pudo concluir que la misma que se pueden tomar formas diferentes de explotación de datos tanto a nivel de formato de registros como a los periodos de tiempos en los cuales se puede distribuir la información.

Un estudio que implementó la minería de datos, fue desarrollado por Ignacio Perversi (2007), su aplicación fue para la exploración y detección de patrones delictivos en Argentina, a raíz de que en los últimos años detectó un aumento en los índices delictivos y los niveles de violencia, por lo cual implementa la herramienta en el análisis de información criminal en Argentina y comprobar su efectividad y valor agregado, se trabajó en la identificación y detección de patrones de homicidios dolosos cometidos en ese país. En el caso de la inteligencia criminal, la gran cantidad de información y de variables que intervienen, justifican el uso de herramientas más potentes que la estadística convencional que permitan determinar relaciones multivalentes subyacentes, demostró que la minería de datos fue factible en su aplicación para obtener la información, el valor agregado fue la generación de conocimiento ya que evaluó con estadísticas toda información obtenida de los últimos años, donde obtuvo una solución razonable para proponer mejoras y obtener conocimiento de lo generado.

1.3 Planteamiento del Problema

Un elemento importante en toda organización es la información, en especial para aquellas compañías que están distribuidas en diferentes lugares ya sea en otras ciudades o países, que manejan muchos clientes, proveedores, artículos, etcétera, y sobretodo que manejan grandes volúmenes de información, por lo cual nace la necesidad de contar con una base de datos para obtener la información deseada en el momento oportuno.

El realizar consultas convencionales en las bases de datos no es suficiente para resolver problemas de negocios, sino que se debe considerar seguir una metodología ordenada que permita aplicar herramientas tecnológicas y técnicas disponibles para obtener conocimiento y resultados que los beneficien. Los negocios necesitan estar a la vanguardia para mantenerse en un ámbito competitivo e identificar estrategias claves para satisfacer las necesidades del mercado a través de una tecnología que les permitan obtener información efectiva y segura desde una base de datos.

En el caso de la agencia automotriz Renault, al presentar un crecimiento repentino, le falta obtener información de manera rápida que apoye en la toma de decisiones, principalmente por no tener un panorama más definido de las ventas surgidas en periodos anteriores limitando a la agencia al momento de optar por estrategias benéficas para el departamento.

El problema consiste en que la agencia automotriz no cuenta con herramientas de minería de datos que le permita obtener información de ciertos periodos que se pudieran consultar en un momento determinado.

1.4 Justificación

El motivo principal de este proyecto es identificar las herramientas y técnicas que ayuden a conocer las ventas e identificar los periodos con mejor demanda para apoyar en la toma de decisiones en el departamento de ventas de la agencia automotriz Renault.

Algunos datos se pueden conocer en su momento, sin embargo, no se puede ir más allá para obtener uno en específico, para lograr esto es necesario contar con una estructura de una base de datos que permitirá la fácil extracción de información.

Con la implementación de una herramienta de minería de datos, se pretende analizar bases de datos con la finalidad de encontrar información relevante para las ventas mediante la utilización de cubos de información que permitan establecer estrategias de ventas y así la agencia podrá implementar mejoras en los procesos de ventas.

1.5 Objetivos

Objetivo General.

Implementar en la agencia automotriz Renault una herramienta de minería de datos que permita obtener información de periodos determinados a través de una metodología aplicada para dar soporte a la toma de decisiones.

Objetivos Específicos.

- Seleccionar y aplicar la metodología que favorezca en la minería de datos para generar información.
- Determinar e identificar los valores de las bases de datos existentes, para la extracción de información.
- Transformar información obtenida en la base de datos para la creación de una nueva estructura de datos
- Realizar comparativo de periodos determinados que se tengan como base para favorecer en la aplicación de estrategias de venta en la agencia automotriz Renault.
- Analizar información obtenida de las ventas.
- Generación de gráficas para el departamento de ventas.

1.6 Alcances

- Investigar y probar una metodología para estructurar un proyecto de minería de datos y obtener una serie de pasos que apoyarán para la obtención de información.
- Implementar técnica en el departamento de ventas, para obtener información de las ventas por periodos
- Obtener un modelo que se basará en la identificación de periodos.
- Establecer si los indicadores de validación son adecuados para determinar si las ventas por periodos son idóneas.

1.7 Limitaciones

- Aplicación de la técnica en el departamento de ventas en la agencia automotriz Renault.
- Dispersión para el proceso por región, la mayor parte de la información se centra en Tijuana, y la parte menor de la misma se ubica en Mexicali y Hermosillo.
- Comparativo de la información de ventas por periodos paralelos en la agencia automotriz Renault.
- Cantidad de información, que se disponen de series continuas y completas en un periodo de seis meses en la agencia automotriz Renault.

CAPÍTULO II.

MARCO TEÓRICO

2.1 Minería de datos

La minería de datos tiene una gama amplia de utilidad para las empresas en la actualidad, a través de técnicas, clasificaciones y predicciones se puede obtener información relevante otorgando un conocimiento amplio trabajando en conjunto de algoritmos se podrá determinar un resultado.

2.1.1 Definición

Según Méndez (2006), la minería de datos es un conjunto de herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información interesante, novedosa y útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones.

Si se analiza la definición anterior, se dice que la minería de datos es un conjunto de herramientas y técnicas, una gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y data warehousing, con técnicas de aprendizaje automático y de estadística, es el descubrimiento eficiente de información valiosa en la Figura 2.1 se puede observar su proceso.

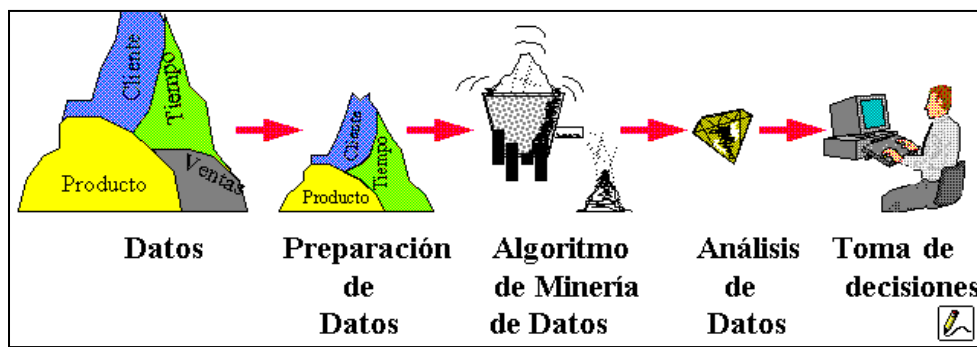


Figura 2. 1 Proceso de la minería de datos (Fuente: Búsqueda de Patrones de Comportamiento en Cubos de Datos, 2006)

Para descubrir conocimiento de la información se pueden utilizar varias formas de análisis por medio de las cuales se puede llegar a identificar patrones y reglas en los datos para luego crear escenarios, esta información se puede representar por medio de modelos matemáticos sobre datos históricos y con esto se crea un modelo de minería de datos. Después de haber creado un modelo de

minería de datos, se puede examinar nueva información a través del modelo para evaluar si se apega a los patrones o reglas definidos.

En el escrito se cuenta con varias definiciones para la minería de datos, algunas de ellas son:

- Witten y Frank (2005), mencionan que la minería de datos es el proceso de descubrir patrones en los datos. Los datos se presentan en grandes cantidades. Los patrones descubiertos deben ser significativos de manera que se permitan ventajas, por lo general, de tipo económicas.
- Sumathi y Sivanandam (2006), indican que la minería de datos es el proceso eficiente, no trivial, de extraer información valiosa (patrones y tendencias) de una gran colección de datos.
- Hand y Smyth (2001), dicen que la minería de datos es el análisis de conjuntos de datos observados, a menudo extensos, para encontrar relaciones insospechadas y resumir los datos en forma comprensible y útil para el usuario de la información.
- Berry y Linoff (2004), menciona que la minería de datos, es la exploración y análisis de grandes cantidades de datos para descubrir reglas y patrones significativos.

Estas cuatro definiciones tienen en común la extracción de información potencialmente útil (patrones, asociaciones o relaciones entre los datos) para los usuarios finales. Algunas otras definiciones consideran además del proceso de descubrir patrones válidos, la exploración y análisis de series extensas de datos.

La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes

heterogéneas, que no tienen el formato adecuado o contienen ruido, para poder interpretar y evaluar los resultados obtenidos.

El proceso completo consta de las siguientes etapas según Moreno, Quintales, García y Polo (2008):

- Selección: identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.
- Preprocesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
- Conversión de datos en un modelo analítico.
- Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos.

Aunque los pasos anteriores se realizan en el orden en que aparecen, el proceso es altamente iterativo, estableciéndose retroalimentación entre los mismos. Además, no todos los pasos requieren el mismo esfuerzo, en general, la etapa de preprocesamiento es la más costosa ya que representa aproximadamente el 60 % del esfuerzo total, mientras que la etapa de minería sólo representa el 10%. (Moreno, Quintales, García y Polo, 2008)

2.1.2 Descripción del proceso de minería

Indican Gilberto Martínez y Adolfo Guzmán (1994), que para realizar la minería se siguen los siguientes pasos:

- Definir el cubo de datos o espacio de búsqueda de mineros.
- Generalización o definición de los niveles de búsqueda en cada una de las dimensiones del cubo.
- Generar los datos y cargar el cubo de datos.
- Definir los horarios de trabajo de los procesos de minería.
- Generar las preguntas (definir región y patrón a buscar).
- Solicitar ejecución del proceso de extracción y análisis.

- Ejecución de la extracción de la región solicitada y la búsqueda del patrón.
- Revisar e Interpretar los resultados.

2.1.3 Definición de cubo de datos.

Un usuario selecciona las variables de sus bases de datos de sus sistemas operacionales, en las que desea buscar patrones o comportamientos de interés. Según el número de variables seleccionadas es la dimensión del cubo formado, a cada variable se le denomina una *dimensión* o un eje del cubo (ver Figura 2.2 y 2.3).

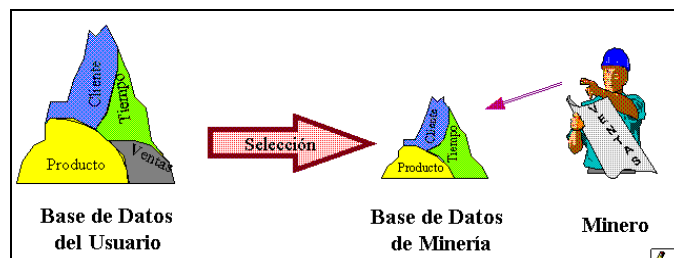


Figura 2. 2. Variables de las bases de datos de sistemas operacionales (Fuente: Búsqueda de Patrones de Comportamiento en Cubos de Datos, 2006)

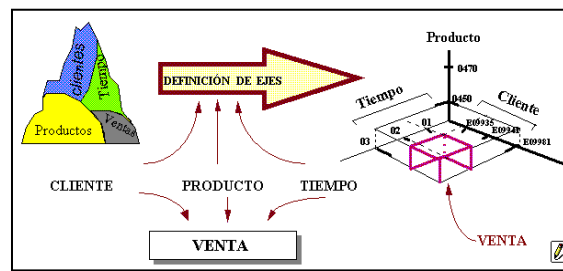


Figura 2. 3. Dimensiones o ejes que forman los cubos y por los cuales los mineros se desplazarán (Fuente: Búsqueda de Patrones de Comportamiento en Cubos de Datos, 2006)

2.1.4 Clasificación

Es una de las principales tareas en el proceso de minería de datos que se emplea para asignar datos a un conjunto predefinido de variables. El objetivo de la clasificación es encontrar algún tipo de relación entre los atributos de entrada y los registros de salida para comprender el comportamiento de los datos, así mediante el conocimiento extraído se puede predecir el valor de un registro

desconocido. Sin embargo, el mayor problema de la clasificación es que muchas veces no es representativo y no proporciona un conocimiento detallado, sólo otorga predicciones. Algunos algoritmos comprendidos en esta tarea son: clasificación bayesiana, árboles de decisión, redes neuronales artificiales, entre otros. (Castillo, 2008)

Regresión

La regresión es el aprendizaje de una función cuyo objetivo es predecir valores de una variable continua a partir de la evolución de otra variable también continua, la cual por lo general es el tiempo (Sumathi y Sivanandam, 2006). En la regresión, la información de salida es un valor numérico continuo o un vector con valores no discretos. Ésta es la principal diferencia respecto a la clasificación donde el valor a predecir es numérico. Si sólo se dispone de una variable definida se trata de un problema de regresión simple, mientras que si se dispone de varias variables se trata de un problema de regresión múltiple.

A esta tarea también se le conoce como: interpolación, donde el valor o valores predichos están en medio de otros; o estimación, cuando se predice valores futuros. (Castillo, 2008).

2.1.5 Áreas relacionadas con la minería de datos

Minería de datos y almacenes de datos (data warehouse)

Indica Neftalí de Jesús Méndez (2006), que con frecuencia los datos que serán minados se extraen del data warehouse de una empresa. Existe un beneficio real si los datos son parte ya de un data warehouse esto es porque el proceso de depuración de datos para un data warehouse y para la minería de datos son similares.

Si la mayor parte de los datos ya han sido limpiados para un data warehouse, es muy probable que no se necesite una limpieza adicional para minar los datos. La base de datos que se va a minar puede ser un subconjunto lógico en

lugar de uno físico del data warehouse, esto sólo si el DBMS del data warehouse soporta los recursos de demanda de minería de datos.

Si el data warehouse no posee soporte de minería de datos entonces se tendrá que separar las base de datos a minar en:

- Geográfica.
- Data Mart.
- Análisis.
- Data mining.
- Data source.
- Data warehose.

El data warehouse no es un requerimiento indispensable para la minería de datos, la configuración de un data warehouse de múltiples datos, resuelve problemas de integridad de datos y una consulta de los datos por medio de un query, puede ser una tarea enorme que puede tomar mucho tiempo y a un costo elevado. Sin embargo para minar los datos de una base de datos operacional o transaccional se puede cargar la información a una base de datos de sólo consulta, este proceso es muy parecido al de un data mart.

2.1.6 Origen de la minería de datos

Las técnicas de minería de datos son el resultado de un largo proceso de investigaciones. Esta evolución se inicia al almacenar la información de organizaciones en las computadoras, continúa con mejoras en el acceso de datos y más recién se han generado tecnologías para que los usuarios naveguen por la información en tiempo real. La minería de datos toma este proceso evolutivo y va más allá del acceso retrospectivo de los datos y navegación para analizar la información para luego mostrar resultados. (Galviz, 2004)

La minería de datos está lista para la aplicación en la comunidad de los negocios ya que ahora cuenta con un soporte de tres tecnologías que la hacen madura:

- Recopilación de datos de forma masiva
- Computadoras poderosas con multiprocesadores
- Los algoritmos de minería de datos

Las bases de datos comerciales crecen a proporciones sin precedentes y conjuntamente la necesidad de motores de búsqueda mejorados.

Los algoritmos de minería de datos personalizan técnicas que han existido por lo menos desde hace 10 años, pero que hasta ahora han podido ser implementadas de tal forma que se consideran herramientas confiables y que consistentemente ejecutan los métodos estadísticos más antiguos.

En la evolución de la forma en que se manejan los datos de los negocios, cada nuevo paso que se construye sobre la base es un previo, por ejemplo, el acceso dinámico de datos es crítico para las aplicaciones de navegación de datos y la habilidad de almacenamiento de grandes bases de datos es crítica para la minería de datos. (Galviz, 2004)

Desde el punto de vista de los usuarios, el manejo de la información ha evolucionado y ha permitido responder nuevas preguntas sobre los negocios de manera rápida y exacta, a continuación se muestran los pasos evolutivos que se han dado:

- Recopilación de datos
- Acceso de datos

Aplicaciones de la minería de datos

La minería de datos es cada vez más popular debido a la contribución substancial que puede hacer. Puede ser usada para controlar costos así como también para contribuir a incrementar las entradas. Muchas organizaciones manejan la minería de datos para ayudar a manejar todas las fases del ciclo vital del cliente, donde incluyen la adquisición de nuevos clientes, para aumentar los ingresos con clientes existentes y mantener bien a la clientela. (Muñoz, 2002)

Determina características de clientes buenos (trazado de perfil), una compañía puede determinar conjuntos con características similares. El perfil de los clientes que hayan comprado un producto en particular, ello puede enfocar la atención en clientes similares que no hayan comprado ese producto (de venta cruzada).

Cerca de perfilar clientes que se han ido, lo que la compañía hace para retener los clientes que están en riesgo de alejarse, porque es normalmente un poco menos caro retener un cliente que conseguir uno nuevo. Las ofertas de minería de datos se valoran a través de una amplia efectividad de industrias. Las compañías de telecomunicaciones y tarjeta de crédito son dos de los conductores para aplicar minería de datos, para detectar el uso fraudulento de sus servicios.

Compañías aseguradoras y bolsas de valores se interesan también al aplicar esta tecnología para reducir el fraude. Las aplicaciones médicas son otra área fructífera; la minería de datos se puede utilizar para predecir la eficiencia de procedimientos quirúrgicos, las pruebas médicas o medicaciones. Las compañías activas en el mercado financiero, usan minería de datos para determinar el mercado y características de industria así como para predecir el comportamiento de las compañías individuales y mejorar el sistema de inventarios. Los minoristas tienen mayor uso de la minería de datos, para decidir qué productos en particular deben mantener en inventario para no abastecerse de productos innecesarios, así como para evaluar la eficacia de promociones y ofertas. (Muñoz, 2002)

Las firmas farmacéuticas poseen grandes bases de datos de los compuestos químicos y de material genético en las cuales hay sustancias que pueden ser muy buenas candidatas para minar, esto con el objetivo de determinar cómo se pueden desarrollar nuevos agentes para los tratamientos de determinadas enfermedades.

Retos de la minería de datos

Nadie duda que la experiencia sea el elemento fundamental del conocimiento y la sabiduría. La asimilación de hechos pasados permite enfrentar

al futuro con más posibilidades de éxito, sin tener que recordar todos los detalles del pasado. Esto es claro en las personas, pero, ¿Cómo puede aplicarse a las corporaciones?

¿Qué promete la minería de datos?, una tecnología informática es infraestructura fundamental de las grandes organizaciones y permite, hoy en día, registrar con lujo de detalle, los elementos de todas las actividades con asombrosa facilidad. La tecnología de bases de datos permite almacenar cada transacción y muchos otros elementos que reflejan la interacción de la organización con todos sus interlocutores, ya sean otras organizaciones, sus clientes, o internamente, sus divisiones, sus empleados, etcétera. Tenemos pues, un registro bastante completo del comportamiento de la organización. Pero, ¿Cómo traducir ese voluminoso conjunto de datos en experiencia, conocimiento y sabiduría corporativa que apoye efectivamente la toma de decisiones, especialmente al nivel gerencial que dirige el destino de las grandes organizaciones? ¿Cómo comprender el fenómeno?, ¿Cómo tomar en cuenta grandes volúmenes de datos? (Thearling, 2012)

La tecnología promete analizar con facilidad grandes volúmenes de datos y reconocer patrones en tiempo y espacio que soportarán la toma de decisiones y construirán un conocimiento corporativo de alto nivel.

Uno de los retos en la minería de datos es que la tecnología parece robusta y lista para su aplicación, dado el gran crecimiento de empresas que comercializan software con diferentes técnicas. Más aún, gran parte de estas técnicas son una combinación directa de madurez en tecnología de bases de datos y "data warehousing", con técnicas de aprendizaje automático y de estadística. Sin embargo, la tecnología enfrenta aún varios retos. (Thearling, 2012)

El primero de estos retos, es la facilidad con que se puede caer en una falsa interpretación; para explicarlo, basta reconocer que las primeras y más maduras técnicas para el análisis de datos, con el fin de modelar un fenómeno, provienen de la estadística. Todos saben que existe la posibilidad de ser engañados por la estadística; no todos tienen un sólido entendimiento de la

matemática, los supuestos y el modelado para entender a la perfección el riesgo o margen de error en un ejercicio de inferencia estadística, pero todos operan y funcionan con resúmenes e indicadores estadísticos generalmente muy simples. Cuando se dice que una gran decisión se basó en la información disponible, típicamente es una serie de promedios y estimadores estadísticos que presentan una generalización de un gran volumen de datos, donde se hace una inferencia. Nótese la facilidad y el poder que proporciona la estadística. Si se retoma el ejemplo del tesorero bancario, la decisión de cerrar agencias o reducir empleados no se basa en una revisión de los miles de datos posibles, pero sí en indicadores (estadísticos), como el valor de la transacción promedio, el salario promedio, etcétera, (Hand, Mannila y Smyth, 2001).

La estadística es una herramienta poderosa y es elemento crucial en el análisis de datos. Sin embargo, a veces se enfrentan problemas muy serios en la interpretación de sus resultados. El ejemplo típico es que, usualmente, no se recuerdan que estos resultados se aplican a grupos (poblaciones) y no a individuos. Estos peligros se ven amplificados en el uso de software de Minería de Datos. Dichas herramientas informáticas pueden poner a disposición de un "analista" o minero de datos, la posibilidad de crear fácilmente indicadores, resúmenes, gráficas, y aparentes tendencias, sin un verdadero entendimiento de lo que se refleja. Es decir, resulta más fácil hacer creíble una falsedad, posiblemente porque la produjo una computadora, con muchas gráficas y con base en muchos datos, eso sí, en un instante. Así que el reto es doble. ¿Cómo hacer las herramientas de minería de datos accesibles a cualquiera, hasta aquel que no sabe lo más mínimo de estadística, pero que sus resultados e interpretaciones sean válidos? (Thearling, 2012)

Es importante que la herramienta tenga un gran elemento de accesibilidad para que su producción sea rentable. Un ejemplo de esto son las bases de datos relacionales, pues su diseño, modelado, y las herramientas alrededor de los manejadores, han hecho posible que no se requiera de una gran especialización para tener una gran cantidad de usuarios y que, por lo tanto, el mercado sea

extenso para mantener a los que producen manejadores de datos, (Hand, Mannila y Smyth, 2001).

Privacidad

Cuando la minería de datos era aún emergente, se llegó a pensar que no presentaba ningún peligro o riesgo para la privacidad de los clientes. Hoy en día, se piensa todo lo contrario, sin embargo, no existe un marco jurídico que haya mantenido el paso con el avance tecnológico. Esto es, hoy en día, las corporaciones comercializan con millones de perfiles personales, sin que aquellos a que se refieren los datos intercambiados, estén en posibilidad de intervenir. Cada llamada telefónica, cada transacción bancaria, cada compra en un supermercado, es registrada en una computadora, y si la compañía de teléfonos, el banco y el club de supermercados combinan sus bases de datos, están en condiciones de elaborar un perfil muy completo. Este perfil definiría a más de una persona (y no como los que están en condiciones de conocer a sus vecinos). Si a esto añadimos qué sitios de WEB visita, qué y dónde se compró con la tarjeta de crédito, etcétera, no existe ninguna privacidad. El problema va desde las definiciones de qué constituye privacidad y quién es el propietario de los datos, hasta qué tanto de un individuo, está en posibilidad real de compilarse. (Hand, Mannila y Smyth, 2001)

2.1.8 DMQL (Lenguaje de minería de datos de consulta)

Las herramientas emergentes de minería de datos y los sistemas llevan naturalmente a la demanda de un lenguaje de minería de datos de consulta poderoso, encima de que muchas interfaces de usuarios interactivas y gráficas pueden desarrollarse. Esto motiva a diseñar un lenguaje de consulta de minería de datos, DMQL, para diferentes tipos de conocimiento en bases de datos de relacionales.

En este sentido, el éxito de los sistemas debe acreditarse en parte a la normalización de lenguajes de consulta. Las actividades de normalización recientes en sistemas de base de datos, tal como el trabajo relacionado con el SQL (Structured Query Language), el OMG (Object Management Group) y el

ODMG (Object Data Management Group), muestra de nuevo la importancia de una base de datos estándar. (Méndez, 2006)

Habitualmente, existen muchas interfaces de usuarios gráficas para diversas tareas de la minería de datos. Sin embargo, es importante comprender los mecanismos fundamentales de los métodos de minería de datos. (Méndez, 2006)

2.1.9 Minería de Datos para la toma de decisiones en las actividades de negocio

Desde un punto de vista más pragmático y asociándolo directamente a las actividades de negocios, la minería de datos es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información no estructurada (interna y externa a la compañía) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento y así dar soporte a la toma de decisiones sobre el negocio.

Marcano Aular y Talavera Pereira (2007), destacan que desde un punto de vista más teórico, la minería de datos se define como el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos, a través de grandes cantidades de datos, al tener de esta manera el encuentro de relaciones o patrones entre los datos procesados.

Por su parte, Marcano y Talavera (2007) explican que los datos tal cual se almacenan en las bases de datos no suelen proporcionar beneficios directos; su valor real reside en la información que podamos extraer de ellos, es decir, información que nos ayude a tomar decisiones o a mejorar la comprensión de los fenómenos que nos rodean. Ejemplos de ello pueden ser: contrastar que todo va bien, analizar diferentes aspectos de la evolución de la empresa, presentar información de forma más intuitiva, comparar información en diferentes períodos de tiempo, comparar resultados con previsiones, identificar comportamientos y

evoluciones excepcionales, confirmar o descubrir tendencias e interrelaciones, entre otras acciones.

Para tal efecto, se deben definir medidas cuantitativas para los patrones obtenidos (precisión, utilidad y beneficio obtenido), para establecer medidas de interés que consideren la validez y simplicidad de los patrones obtenidos mediante alguna de las técnicas de Minería de Datos. El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados, o simplemente registrar la información conseguida y suministrarla a quien esté interesado.

Cabe resaltar que los efectos sobre la eficiencia de los resultados en la minería de datos como apoyo a la toma de decisiones donde se generan miles de opiniones desde diferentes perspectivas, entre ellas se puede destacar: la imposibilidad de encontrar conclusiones únicas referentes a la evaluación de un resultado obtenido con esta técnica. Sin embargo, la consecuencia de los datos arrojados es interesante, por algo más que su precisión.

2.2 Metodologías de Minería de Datos

Al momento principales metodologías a utilizar en la realización de proyectos de minería de datos son:

2.2.1 CRISP-DM (Cross Industry Standard Process for Data Mining)

Definición.

Menciona Gallardo (2003), es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, sus orígenes se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimler-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases) [Reinartz, 1995], [Adraans, 1996], [Brachman, 1996], [Fayyad, 1996], el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining).

Surge de la necesidad de tres empresas interesadas en realizar proyectos de minería de datos, esta proporciona un vistazo general del ciclo de vida de un proyecto de minería de datos. Cada fase se divide en un conjunto de tareas, relacionadas entre sí, esas tareas a su vez se componen de actividades específicas y de un conjunto de resultados concretos.

Esta muestra el camino que se debe seguir en el proceso de extracción del conocimiento y permite determinar las actividades a desarrollar en cada etapa, de forma que se puedan lograr los objetivos propuestos.

Los secretos del éxito de proyectos de minería de datos radica en que tan bien se han seguido o desarrollado estas etapas. Lo ideal es evitar el rompimiento del ciclo de la minería de datos, al entender las distintas formas en que se puede fallar y que acciones preventivas tomar para evitar estas posibles fallas.

Es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining. Está dividida en 4 niveles de abstracción como se muestra Figura 2.4 están organizados de forma jerárquica en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Data Mining, en una serie de seis fases. (Gallardo, 2003)

El diagrama ilustra el ciclo de vida de un proyecto de ciencia de datos, organizado en un círculo con los siguientes componentes:

- Comprensión del negocio** (rectángulo naranja superior izquierdo)
- Análisis de los datos** (rectángulo naranja superior derecho)
- Preparación de los datos** (rectángulo naranja inferior derecho)
- Modelamiento** (rectángulo naranja inferior izquierdo)
- Evaluación** (rectángulo naranja inferior izquierdo, debajo de Modelamiento)
- Despliegue** (rectángulo naranja superior izquierdo, a la izquierda de Comprensión del negocio)

En el centro del ciclo se encuentra un cilindro azul etiquetado como **Datos**.

Las flechas azules indican las relaciones entre los pasos:

- Una flecha horizontal bidireccional conecta **Comprensión del negocio** y **Análisis de los datos**.
- Una flecha curva desciende desde **Análisis de los datos** hacia **Preparación de los datos**.
- Una flecha curva sube desde **Preparación de los datos** hacia **Modelamiento**.
- Una flecha horizontal conecta **Modelamiento** y **Evaluación**.
- Una flecha curva sube desde **Evaluación** hacia **Comprensión del negocio**.
- Una flecha curva desciende desde **Evaluación** hacia **Despliegue**.
- Una flecha curva conecta **Despliegue** y **Comprensión del negocio** a lo largo del borde exterior del ciclo.

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que

deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

El estándar incluye un modelo y una guía, estructurados en **seis fases** menciona Chapman, Clinton, Kerber (1999-2000), que algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores:

Comprensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica).

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito).
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio,...).
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito).
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas).

Comprensión de los datos (Familiarizarse con los datos y tener presente los objetivos del negocio).

- Recopilación inicial de datos.
- Descripción de los datos.
- Exploración de los datos.
- Verificación de calidad de datos.

Preparación de los datos (Obtener la vista minable o dataset).

- Selección de los datos.
- Limpieza de datos.
- Construcción de datos.
- Integración de datos.
- Formateo de datos.

Modelado (Aplicar las técnicas de minería de datos a los dataset).

- Selección de la técnica de modelado.
- Diseño de la evaluación.
- Construcción del modelo.
- Evaluación del modelo.

Evaluación (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio).

- Evaluación de resultados.
- Revisar el proceso.
- Establecimiento de los siguientes pasos o acciones.

Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización).

- Planificación de despliegue.
- Planificación de la monitorización y del mantenimiento.
- Generación de informe final y Revisión del proyecto.

El CRISP-DM no ha sido construido a una manera teórica, académica que trabaja de principios técnicos sino a un ámbito más práctico. Ambos de estos accesos a metodologías que se desarrollan han sido intentados en el pasado, pero raras veces conducían a lo práctico, lo acertado, y extensamente ha adoptado normas. Tiene éxito porque está profundamente basado en la experiencia práctica, la experiencia del mundo real de cómo la gente conduce proyectos de minería de datos con sus esfuerzos e ideas en todas partes del proyecto. (Pino, 2000)

En la Figura 2.5 se observan claramente las fases que componen la metodología CRISP-DM en el orden secuencial natural de las mismas.



Figura 2.5, Fases componentes de la metodología CRISP-DM. (Fuente: Pino, 2000)

Cada una de estas fases se divide en distintas fases de nivel inferior que indican tareas generales a realizar dentro de la misma. A su vez, estas tareas de segundo nivel, son divididas en tareas específicas donde se describen las acciones que deben ser desarrolladas en situaciones específicas. Por ejemplo, si entre las tareas generales se encuentra “limpieza de datos” en el tercer nivel se indican las tareas que deben desarrollarse para un caso específico como por ejemplo “limpieza de datos numéricos” o “limpieza de datos categóricos”. En un cuarto nivel se recogen acciones, decisiones y resultados sobre el proyecto de Explotación de Información. Esta abstracción de procesos puede verse gráficamente en la Figura 2.6. (Pino, 2000)

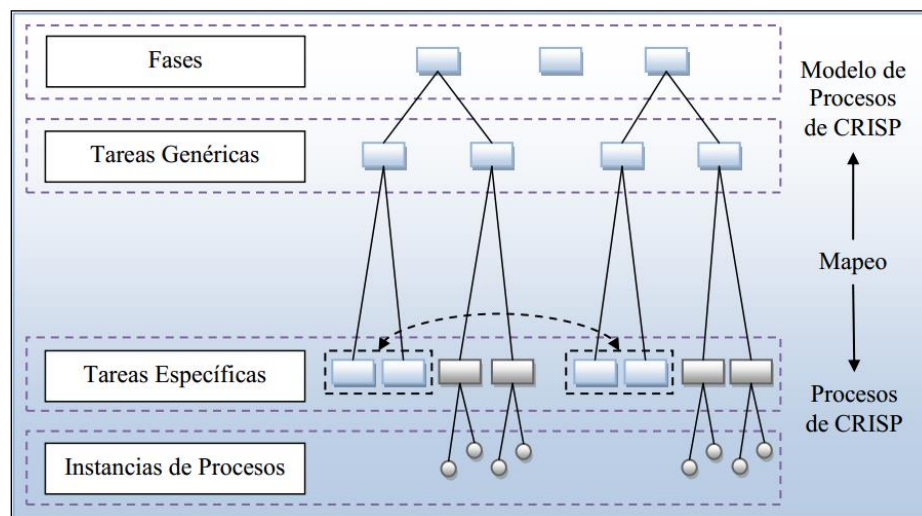


Figura 2.6 Niveles de abstracción de procesos de CRISP-DM (Fuente: Pino, 2000)

2.2.2 SEMMA (Sample, Explore, Modify, Model and Assess)

Definición

La metodología Semma es una técnica de minería de datos, que permite la selección, exploración, y el modelado de grandes cantidades de información, con el fin de descubrir nuevas oportunidades de negocio en la exploración de los datos. (Pérez y Santin, 2007). Está definida por el nombre de las fases que la componen, muestra Figura 2.7:

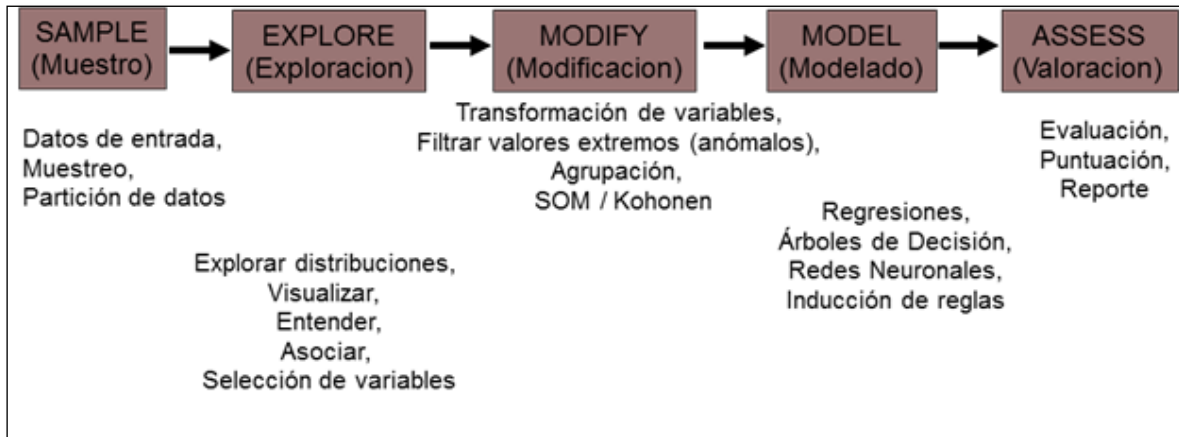


Figura 2.7 Fases de metodología Semma (Fuente: Pérez, González, 2007)

El desarrollo de las fases de la metodología SEMMA, menciona la necesidad de la comprensión del negocio como punto central para el desarrollo de la investigación de cualquier proyecto.

Con esta metodología se ayuda, de manera mucho más sencilla, con la realización de exploración estadística y las técnicas de visualización, además de seleccionar y transformar las variables predictivas, a la vez de novelizar las variables para predecir resultados para llevar a la fiabilidad de un modelo.

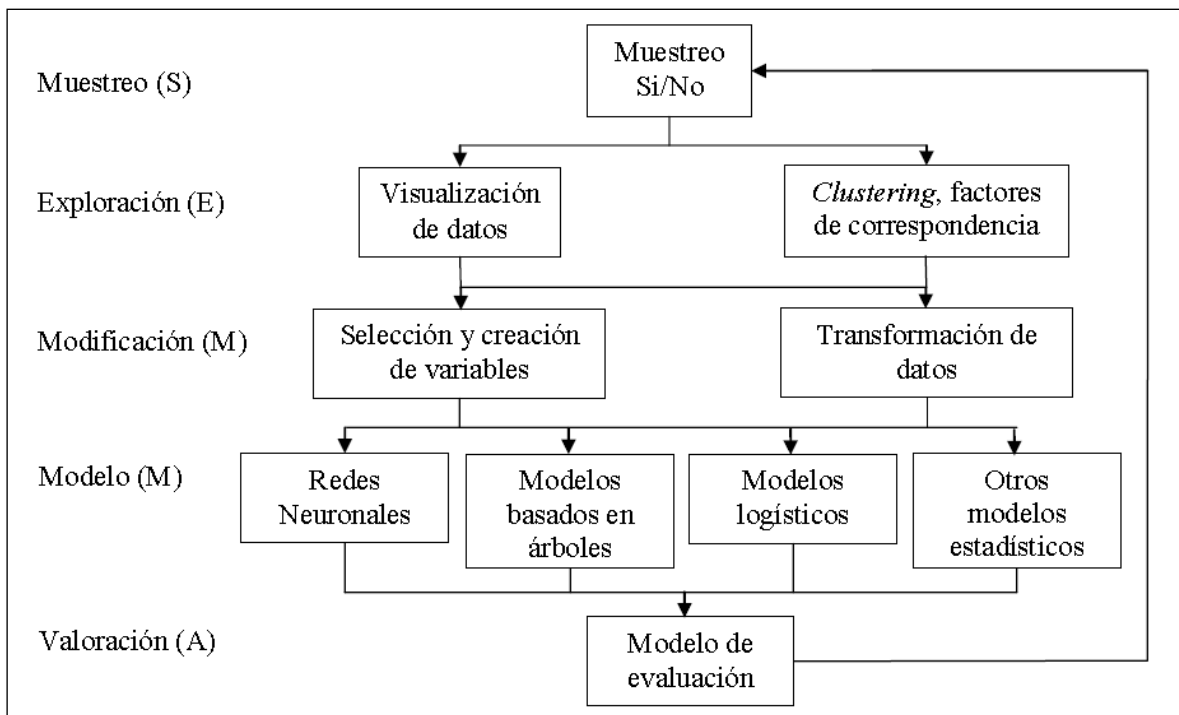


Figura 2.8 Dinámica de la Metodología SEMMA. Extraído de (Fuente: Vendrell, 2011)

El desarrollo de la metodología SEMMA menciona Juan Ángel Vendrell (2011) está conformado por cinco fases, mostradas en la Figura 2.8, las cuales se describirán a continuación:

1. SAMPLE:

Este proceso comienza con una extracción de la población muestral sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método de muestreo se denomina muestreo aleatorio simple. (Vanrell, 2011)

2. EXPLORE:

Se procede a una exploración de la información disponible con el fin de simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo. (Vanrell, 2011)

3. MODIFY:

Consiste en la manipulación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo. Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. (Vanrell, 2011)

4. MODEL:

Para poder hacer el modelo se utilizan técnicas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy, árboles de decisión, reglas de asociación y computación evolutiva. (Vanrell, 2011)

5. ASSES:

Esta fase consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones muestrales. (Vanrell, 2011)

2.2.3 Diferencias entre SEMMA - CRISP-DM

Tanto la metodología SEMMA como la CRISP-DM desarrollan el proyecto Data Mining en distintas fases que se encuentran interrelacionadas para producir un desarrollo iterativo e interactivo.

Una gran diferencia se encuentra en el comienzo del proyecto que será analizado ya que para la metodología SEMMA se centra más en las características técnicas en cada proceso, donde empieza con el muestreo de los datos; en cambio la metodología CRISP-DM realiza la minería de datos, de manera más amplia de acuerdo a los objetivos empresariales y el análisis del problema empresarial para su transformación en un problema técnico. (Pérez López y Santin Gonzalez, 2007)

De manera global se puede considerar que la metodología CRISP-DM es más cercana al concepto real de cada proyecto, donde se integrara con una metodología de gestión de proyectos específica que completaría las tareas administrativas y técnicas. También se encuentra una gran diferencia en el tema comercial. Donde SEMMA es ligada a los productos SAS, permitiéndole ser abierta a sus aspectos generales, donde se encuentra implementada. Por otro lado la metodología CRISP-DM ha sido diseñada como metodología neutra, con distribución libre y gratuita.

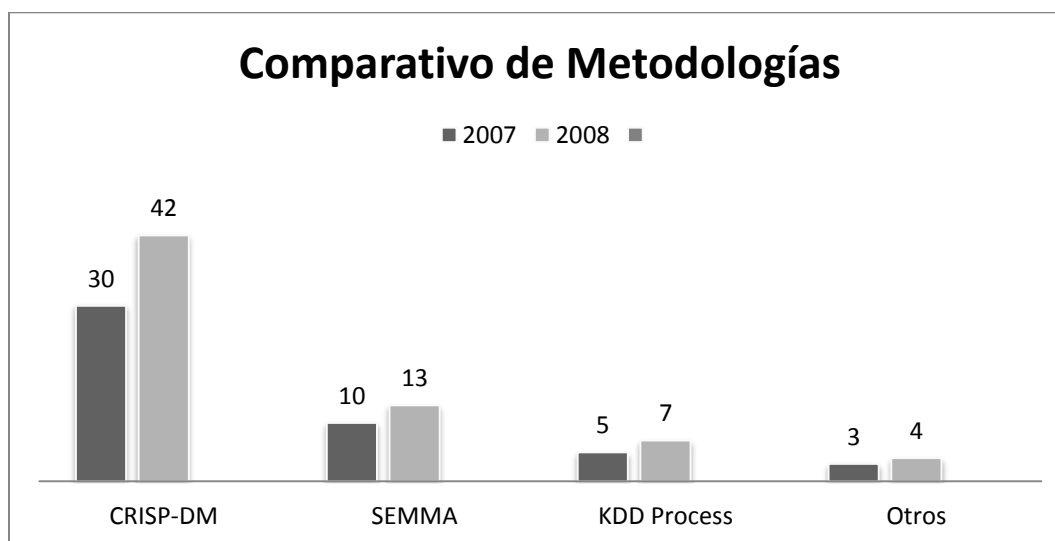


Figura 2.9 Gráfica comparativa (Fuente: Pérez López y Santin González, 2007)

Las metodologías SEMMA, CRISP-DM esencialmente son muy parecidas ya que están compuestas por etapas o fases que interactúan entre sí, en la Figura 2.9 se muestra un comparativo de años anteriores de estas técnicas.

Es posible resaltar que la metodología CRISP-DM es más completa debido a que tiene toda una fase dedicada al entendimiento del negocio.

La Tabla 2.1 muestra un cuadro comparativo entre las diferentes metodologías descritas hasta aquí.

Diferencia entre SEMMA y CRISP-DM		
Actividades	SEMMA	CRISP-DM
Permite elección libre de las herramientas	NO	SI
Cantidad de fases	5	6
Todas las fases pueden relacionarse	NO	SI
Considera los motivos del proyecto	NO	NO
Considera la naturaleza del interés de las partes	NO	NO
Considera otros aspectos no técnicos	NO	SI
Identifica claramente las variables sobre las cuales el proyecto tiene impacto	NO	NO
Esta detalla paso a paso cada etapa del método	NO	NO
Identifica problemas de inteligencia de negocio (PIN)	NO	SI
Identifica una caracterización abstracta de negocio PIN	Parcialmente	NO
Identifica técnicas de explotación de información (TEI) utilizables	SI	SI
Identifica relaciones entre las TEI y los PIN	Parcialmente	NO
Identifica procesos de explotación de información (procesos PINxTEI)	Parcialmente	NO

Tabla 2. 1 Comparativo de metodologías (Fuente Propia)

Indica Gallardo (2003), en referencia a la tecnología SEMMA está más ligada a los aspectos técnicos de la explotación de datos. En cuanto a la otra, está más enfocada en el negocio en sí; es decir en la aplicación de la Minería de Datos a los negocios. Esta diferencia se ve específicamente en la primera etapa donde SEMMA arranca directamente en el trabajo de datos mientras que CRISP-DM empiezan por el estudio del negocio y sus objetivos, y luego recién se transforma en un problema técnico.

Se puede decir que uno de los motivos por los cuales se desea escoger para el presente proyecto la metodología CRISP-DM es que este sistema está diseñado como una metodología independiente de la herramienta tecnológica a utilizar en la explotación de datos. Esto último la hace más flexible. Otro punto importante es que es de libre acceso y cumple con la característica de ser orientada al negocio. Para esta implementación su desarrollo será aplicado a los datos de la Industria Automotriz.

Desde un punto de vista empresarial la minería de datos puede ser definida como un conjunto de áreas que tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisiones, realmente su finalidad es reunir las ventajas de varias áreas como la estadística, inteligencia artificial, computación gráfica, entre otras, al ser la materia prima base de datos, por tal motivo la importancia de esta herramienta, que trabaja en conjunto con técnicas para poder apoyar a las organizaciones a tomar decisiones y a su vez contar con estrategias para favorecer sus procesos.

2.3 Herramienta de Minería de Datos.

2.3.1 Definición

RapidMiner (anteriormente, YALE, *Yet Another Learning Environment*) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales. En una encuesta realizada por *KDnuggets*, un periódico de minería de datos, RapidMiner ocupó el segundo lugar en herramientas de analítica y de minería de datos utilizadas para proyectos reales en 2009¹ y fue el primero en 2010. (Berry y Linoff, 2004)



La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluye los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en Weka.

2.3.2 Características de RapidMiner

- Desarrollado en Java.
- Multiplataforma.
- Representación interna de los procesos de análisis de datos en archivos XML.
- Permite el desarrollo de programas a través de un lenguaje de script.

Puede usarse de diversas maneras:

- A través de un GUI.
- En línea de comandos.
- En *batch* (lotes).
- Desde otros programas a través de llamadas a sus bibliotecas.
- Extensible.
- Incluye Gráficas y herramientas de visualización de datos.
- Dispone de un módulo de integración con R.

En RapidMiner los proyectos se trabajan mediante una definición de procesos, una definición de procesos es un conjunto de operadores o módulos encadenados en una secuencia. Una definición de proceso es un archivo que se almacena en un repositorio. Un repositorio además de almacenar archivos de definición de procesos, puede almacenar datos precargados en archivos de lectura rápida. (Pérez y Santin, 2007).

2.4 Toma de decisiones

2.4.1 Definición.

La toma de decisiones se define como la selección de un curso de acción entre varias opciones, también puede ser la selección racional de un curso de acción (Hillermann, 2000), las 5 palabras mencionadas a continuación son clave para la toma de decisiones:

- ***¿Qué?***
- ***¿Cuándo?***
- ***¿Dónde?***
- ***¿Quién?***
- ***¿Cómo?***

La importancia de la toma de decisiones es una función que trasciende de la empresa: funcionarios de gobierno no tendrán una mentalidad gerencial o se han preparado para administrar, estas se toman continuamente, por ejemplo, por los jefes de familia o un simple ciudadano.

Según Daft (1997), la toma de decisiones organizacionales se define formalmente como el proceso de identificar y resolver problemas. El proceso contiene dos etapas principales. En la fase de identificación del problema se vigila la información sobre las condiciones ambientales y organizacionales, para determinar si el desempeño es satisfactorio y para diagnosticar la causa de las insuficiencias. La solución del problema tiene lugar cuando se consideran cursos alternativos de información y se selecciona e implanta una opción.

Las decisiones organizacionales varían en complejidad y pueden dividirse en programadas o no programadas. Las *decisiones programadas* son repetitivas, bien definidas y existen procedimientos para resolver el problema. Están bien estructuradas a causa de que los criterios de desempeño suelen estar claros, existe una buena información sobre el desempeño actual, las opciones se especifican con facilidad y hay certeza relativa de que la opción escogida tendrá

éxito. Las decisiones programadas tienen reglas de decisión como cuando hay que reemplazar cierto tipo de equipo por su vida de uso o tiempo de depreciación, como vehículos utilitarios y equipo de cómputo. (Rao, 1996)

Los gerentes y organizaciones de hoy se enfrentan a un mayor porcentaje de decisiones no programadas a causa del entorno de negocios en rápido cambio.

2.4.2 La toma de decisiones y la organización del trabajo en la empresa

Indica Mari (2009), que hoy en día realizar bien el trabajo no es suficiente para el éxito de los proyectos empresariales por lo que no podemos olvidarnos de la influencia de la "inteligencia emocional"; es decir, de adoptar una actitud abierta y responsable de cooperación y adaptarla a las normas y modos de trabajar de nuestra entidad.

La manera más adecuada de proceder a la hora de tomar decisiones en una compañía es analizar la situación desde todos los puntos de vista posibles. Después es necesario actuar con seguridad y convicción. Independientemente también de cuál sea nuestra posición en la empresa, podemos solicitar la cooperación de los demás, y agradecer sus ideas.

Hay que saber lo que puede rendir cada empleado, cuáles son las obligaciones que tiene y exigirle que las cumpla, pero siempre con educación y asertividad, conductas que indican seguridad en nuestra capacidad de liderazgo y confianza en las competencias de los demás.

Tanto la toma de decisiones como la organización laboral se asientan en la facultad de *delegar*. No se trata de dar más trabajo o responsabilidad a los subordinados, sino mejor trabajo, con el fin de que todos nos centremos en lo que es valioso para la empresa. La delegación evidencia el éxito directivo si va acompañada de un buen criterio de selección, formación y refuerzo.

Independientemente del sistema de funcionamiento que adoptemos, es importante que en nuestra organización los objetivos empresariales y departamentales estén bien concretados y definidos, con una doble finalidad:

- Optimizar los recursos disponibles.
- Solucionar, de una manera integrada y eficaz, los problemas reales que puedan generarse.

Como indica Mari (2009), dado el derecho que tenemos todas las personas a operar subjetivamente sin ser enjuiciadas, debe incidirse en la *necesidad de la comunicación* entre todos los interlocutores para la consecución de los objetivos establecidos, con mayor motivo en empresas del ramo de la transmisión de la información. Todo ello acompañado de una eficaz coordinación y de un riguroso método de procedimiento de trabajo, que refuerzan notablemente la implicación y el compromiso del personal para cumplir con los plazos acordados.

Como se señala en la Figura 2.10, el ambiente actual ha incrementado la cantidad y la complejidad de las decisiones y ha creado la necesidad de nuevos procesos de toma de decisiones. Un ejemplo al respecto es que debido a la globalización la tendencia es a trasladar la producción a países de bajos salarios, es una decisión común de los ejecutivos de empresas estadounidenses luchar con la ética de las decisiones en relación con las condiciones de trabajo del 3er mundo y la pérdida de empleos en las pequeñas comunidades de los Estados Unidos. (Rodríguez, 2007)

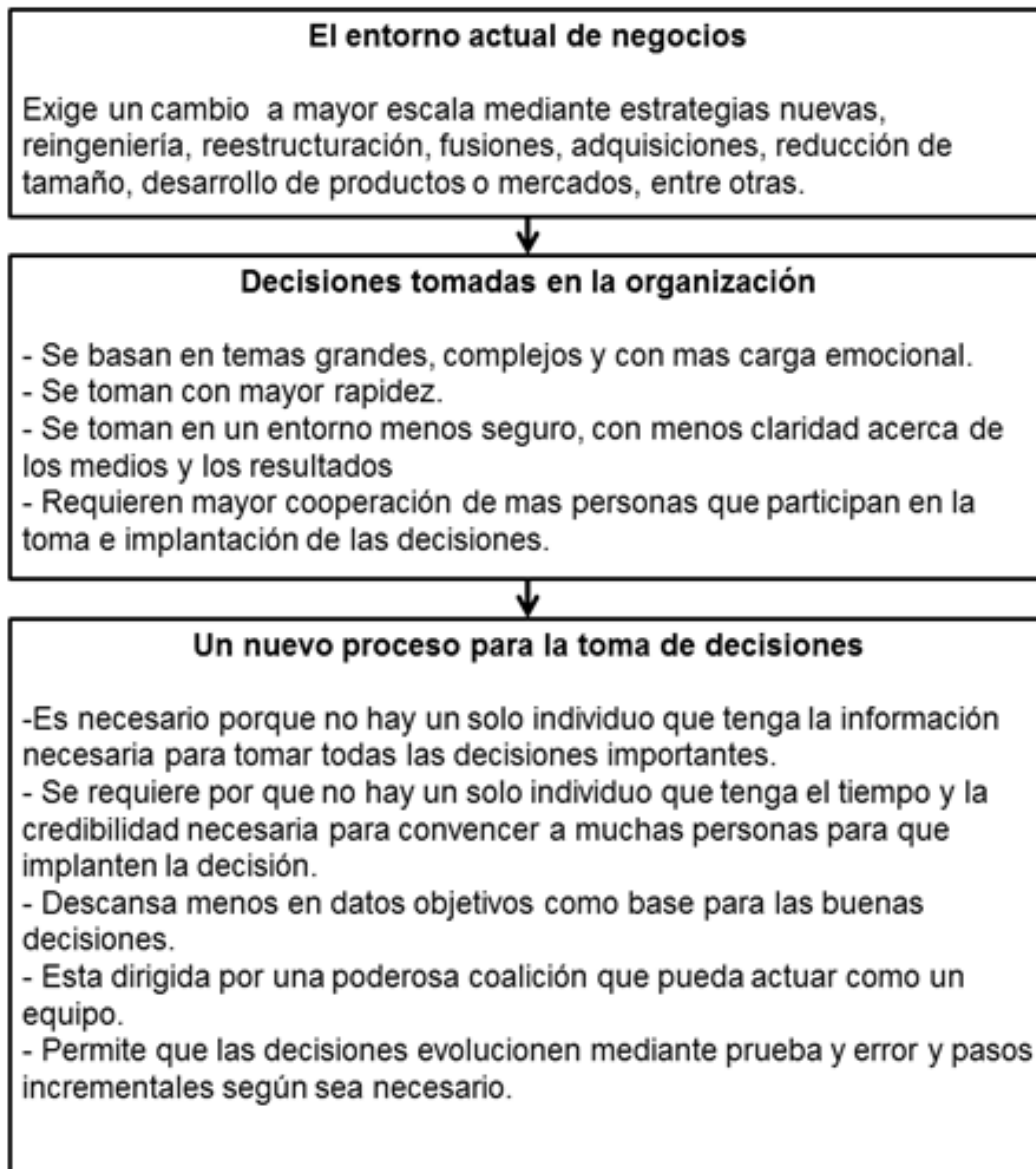


Figura 2.10 Toma de decisiones en el entorno actual (Fuente: Propia)

La toma de decisiones, para ser adecuadamente planteada, exige, según Valda (2009):

- Saber cuáles son los objetivos o la meta a alcanzar.
- Qué camino hay que recorrer, es decir que hay que hacer para llegar a ese objetivo.
- Conocer dos o más alternativas.
- Ser consciente de los límites o dificultades que existen en cada una de ellas.

- Tener la información suficiente.
- Analizar y evaluar estas diferentes alternativas.
- Concluir cuál es la mejor o la más aconsejable, dadas las circunstancias, de esas alternativas posibles.

2.4.3 Etapas en el proceso de toma de decisiones

Por lo general, hemos definido la toma de decisiones como la “selección entre alternativas. Esta manera de considerar la toma de decisiones es bastante simplista, porque la toma de decisiones es un proceso en lugar de un simple acto de escoger entre diferentes alternativas. El proceso de toma de decisiones contiene una serie de ocho pasos que comienza con la identificación del problema, los pasos para seleccionar una alternativa que pueda resolver el problema, y concluyen con la evaluación de la eficacia de la decisión. Este proceso se puede aplicar tanto a sus decisiones personales como a una acción de una empresa, a su vez también se puede aplicar tanto a decisiones individuales como grupales. Vamos a estudiar con mayor profundidad las diversas etapas a seguir para un mayor conocimiento del proceso: (Valda, 2009)

Etapas 1.- La identificación de un problema

El proceso de toma de decisiones comienza con un problema, es decir, la discrepancia entre un estado actual de cosas y un estado que se desea. Ahora bien, antes que se pueda caracterizar alguna cosa como un problema los administradores tienen que ser conscientes de las discrepancias, estar bajo presión para que se tomen acciones y tener los recursos necesarios. Los administradores pueden percibir que tienen una discrepancia por comparación entre el estado actual de cosas y alguna norma, norma que puede ser el desempeño pasado, metas fijadas con anterioridad o el desempeño de alguna otra unidad dentro de la organización o en otras organizaciones. Además, debe existir algún tipo de presión en esta discrepancia ya que si no el problema se puede posponer hasta algún tiempo en el futuro. Así, para iniciar el proceso de decisión, el problema debe ejercer algún tipo de presión sobre el administrador para que

éste actúe. Esta presión puede incluir políticas de la organización, fechas límites, crisis financieras, una próxima evaluación del desempeño etc.

Por último, es poco factible que los administradores califiquen un evento como problema sino tienen la autoridad, dinero, información, u otros recursos necesarios para poder actuar.

Paso 2.- La identificación de los criterios para la toma de decisiones.

Una vez que se conoce la existencia del problema, se deben identificar los criterios de decisión que serán relevantes para la resolución del problema. Cada persona que toma decisiones suele tener unos criterios que los guían en su decisión. Este paso nos indica que son tan importantes los criterios que se identifican como los que no; ya que un criterio que no se identifica se considerará irrelevante por el tomador de decisiones.

Paso 3.- La asignación de ponderaciones a los criterios.

Los criterios seleccionados en la fase anterior no tienen todos la misma importancia, por tanto, es necesario ponderar las variables que se incluyen en la lista en el paso anterior, a fin de darles la prioridad correcta en la decisión. Este paso lo puede llevar a cabo dándole el mayor valor al criterio preferente y luego comparar los demás para valorarlos en relación al preferente.

Paso 4.- El desarrollo de alternativas.

Este paso consiste en la obtención de todas las alternativas viables que puedan tener éxito para la resolución del problema.

Paso 5.- Análisis de las alternativas.

Una vez que se han desarrollado las alternativas el tomador de decisiones debe analizarlas cuidadosamente. Las fortalezas y debilidades se vuelven evidentes según se les compare con los criterios y valores establecidos en los pasos 2 y 3. Se evalúa cada alternativa comparándola con los criterios. Algunas valoraciones pueden lograrse en una forma relativamente objetiva, pero, sin

embargo, suele existir algo de subjetividad, por lo que la mayoría de las decisiones suelen contener juicios.

Paso 6.- Selección de una alternativa.

Este paso consiste en seleccionar la mejor alternativa de todas las valoradas.

Paso 7.- La implantación de la alternativa.

Mientras que el proceso de selección queda completado con el paso anterior, sin embargo, la decisión puede fallar si no se lleva a cabo correctamente. Este paso intenta que la decisión se lleve a cabo, e incluye dar a conocer la decisión a las personas afectadas y lograr que se comprometan con la misma. Si las personas que tienen que ejecutar una decisión participan en el proceso, es más fácil que apoyen con entusiasmo la misma.

Estas decisiones se llevan a cabo por medio de una planificación, organización y dirección efectivas.

Paso 8.- La evaluación de la efectividad de la decisión.

Este último paso juzga el proceso el resultado de la toma de decisiones para ver si se ha corregido el problema. Si como resultado de esta evaluación se encuentra que todavía existe el problema tendrá que hacer el estudio de lo que se hizo mal. Las respuestas a estas preguntas nos pueden llevar de regreso a uno de los primeros pasos e inclusive al primer paso.

Para adoptar algunos tipos de decisiones se suelen utilizar modelos. Estos se pueden definir como una representación simplificada de una parte de la realidad, y ello porque en muchos casos la realidad es tan compleja que, para comprenderla hay que simplificar y tomar de ella los aspectos que resultan más relevantes para el análisis de que se trate y no tener en cuenta los que resultan accesorios. El principal objetivo de un modelo es permitir una mejor comprensión y descripción de la parte de la realidad que representa. Esa mejor comprensión de la realidad permite tomar mejores decisiones. Los modelos se pueden clasificar en atender a numerosos criterios entre ellos:

1. Modelos objetivos y subjetivos. En ocasiones los sucesos no se pueden experimentar objetivamente, y no existen métodos formales para su estudio, por lo que los modelos han de ser informales, subjetivos y basarse en la intuición.
2. Modelos analíticos y de simulación. Los modelos analíticos son aquellos que sirven para obtener soluciones, por lo tanto han de ser resueltos. Los modelos de simulación son representaciones simplificadas de la realidad sobre las que se opera para estudiar los efectos de las distintas alternativas de actuación.
3. Modelos estáticos y dinámicos. Los modelos estáticos son aquellos que no utilizan la variable tiempo, en tanto que los dinámicos son aquellos que incorporan el tiempo como variable o como parámetro fundamental.
4. Modelos deterministas y probabilísticos. En los modelos deterministas se suponen conocidos con certeza todos los datos de la realidad que representan. Si uno o varios datos se conocen sólo en términos de probabilidades, el modelo se denomina probabilístico, aleatorio o estocástico.

2.4.4 Sistema de aprendizaje

Es un programa de computadora que toma decisiones basadas en experiencias acumuladas de casos resueltos con éxito. A diferencia de un sistema experto que resuelve problemas al usar un ordenador pero con un modelo de razonamiento humano, un sistema de aprendizaje puede usar diferentes técnicas para aprovechar el poder computacional de un ordenador muestra figura 2.11, sin relación con el proceso de pensamiento humano. (Páez, 2003)

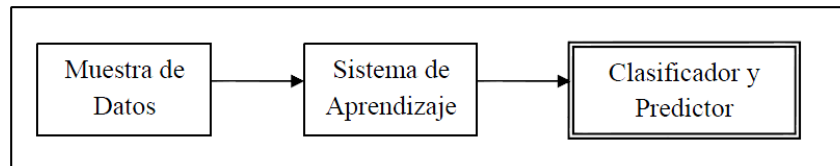


Figura 2.11 Sistema de Aprendizaje (Fuente: Propia)

Una de las tareas más destacadas y básicas del aprendizaje es la de clasificación o predicción. En este caso el sistema tiene disponible un conjunto finito o muestra de ejemplos de casos resueltos. Los datos de cada caso estriban en un conjunto de observaciones o características de ese caso (patrón) y la correspondiente clasificación correcta.

El *aprendizaje* consiste en elegir una estructura modelo y adaptar correctamente unos parámetros de ésta para una óptima clasificación de los datos conocidos. El resultado es un clasificador, sistema que proporciona una decisión para cada patrón admisible que se le proporcione. (Valda, 2009).

2.5 Agencia Automotriz

2.5.1 Definición.

Es una edificación que se encarga de comercializar y dar mantenimiento a los vehículos. En un edificio comercial, el objetivo es tener utilidades. En el caso de las agencias automotrices, la competencia entre marcas es tan grande, que cambiaron constantemente sus productos. El público de México es diferente al de Estados Unidos, Europa o Japón. Por lo tanto el automóvil se compra de una manera distinta, (Noriega, 2011).

La industria automotriz es un sector dinámico que requiere una solución de gestión empresarial que le facilite a los negocios ser más competitivos y redituables. Para hacer frente a las necesidades del mercado, resulta imprescindible contar con información oportuna que facilite el desarrollo de mejores estrategias, basadas en el conocimiento detallado de sus clientes y sus necesidades, (Intelesis, 2007).

2.5.2 Guía de Operación.

Montiel (2006), menciona que cada vez que un cliente visita el departamento de ventas, se lleva una impresión de sus operaciones, ya sea positiva o negativa. No hay términos medios, hacer un trabajo “adecuado” no es

suficiente. Es muy importante que se aproveche al máximo cada visita. Y las visitas iniciales de un cliente son las más importantes. Las investigaciones muestran que su impresión general se forma por sus experiencias en estas cuatro situaciones:

- Preparación / entrega de una unidad nueva.
- Servicio de las millas (u otra revisión).
- Visitas de garantía.
- Primera visita post- garantía.

2.5.3 Preparación pre-entrega de una unidad nueva

No es necesario decir que cada unidad nueva debe ser preparada correctamente, eso significa también sus accesorios. Que su encargado de la preparación use una lista de preparación y pre-entrega y le remita una copia para el cliente.

Se logra el mayor beneficio de la primera experiencia iniciadora del cliente para desarrollar una parte activa en la entrega de la unidad nueva. Se puede asegurar que la primera visita salga bien al tomar en cuenta estos puntos:

- Entregue al cliente una copia de la lista de revisión de preparación y pre-entrega, y coméntela con él.
- Describa el servicio de las millas (u otro intermedio) y la póliza de garantía.
- Hable con el cliente sobre el departamento de servicio: horario, citas, formas de pago, servicios especiales, etc.
- Pida al cliente que llame si tiene dudas, problemas y demás.
- Entregue al cliente su tarjeta y dígame que espera la oportunidad de servirle.

2.5.4 Servicio de las millas u otra “revisión”

La primera visita de servicio debe ser una experiencia fácil y rápida para el cliente. Revisión en general del vehículo además de los puntos específicos cubiertos por los servicios de millas. Luego, se comenta con el cliente lo que se encontró como resultado de la inspección. Este servicio de revisión presenta una

gran oportunidad de ventas. El cliente ya tuvo tiempo de conocer su unidad y ver que tal sirve. Es importante dedicar tiempo a determinar cuáles accesorios le servirán de clientes. El cliente tendrá algunas ideas y puede dirigirlos correctamente.

2.5.5 Visita de garantía

Más clientes se pierden en las visitas de garantía que en ninguna otra clase de las visitas indicadoras. La mayoría de los clientes consideran las visitas de garantía como la prueba de la honestidad de la agencia. Su experiencia de garantía les dice si la agencia en realidad se interesa por satisfacer a los clientes o sólo quiere vender autos.

Estas tres estrategias sirven para que la visita de garantía haga que el cliente se quede en lugar de irse.

2.5.6 Primera visita post garantía.

La primera visita pagada por el cliente también es crítica. Él sabe que su auto ya no está cubierto por la garantía. Debe tomar la decisión de quien le dará el servicio a su auto en el futuro, se revisará para el cliente de una forma objetiva. (Montiel, 2006).

Todo debe salir a la perfección, se debe tener el expediente de servicio del cliente para su revisión, tener cuidado de no “vender de más” ni forzar al cliente a realizar servicio que no deseen realmente.

Se debe tomar en cuenta algunos puntos como sugerencias, mencionados a continuación:

- Contactar al cliente si el contrato de trabajo necesita algún cambio.
- Probar la unidad y revise la calidad del trabajo usted mismo.
- Sin importar que más tenga que hacer, realizar el trabajo a tiempo y dentro del presupuesto.
- Tratar de encontrar algún trabajo menor que se pueda realizar sin cargo, como instalar un foco, por ejemplo.

CAPITULO III.

METODOLOGÍA

En el Capítulo II se dieron a conocer las metodologías existentes para la implementación de la minería de datos, donde se realizó un comparativo de las existentes y su proceso, para este proyecto se considera apropiado utilizar la metodología CRISP-DM, a continuación se muestran las razones de su elección:

- Está enfocada en apoyo del negocio, con apoyo de lo técnico.
- Entendible para el usuario, al trabajar en conjunto con herramientas de fácil entendimiento.
- Sus fases comprenden lo siguiente:
 - Determinación de objetivos y situación actual de la agencia.
 - Preparación de los datos.
 - Construcción de modelos.
 - Validación y resultados.

La metodología mencionada trabajará en conjunto con la herramienta *RapidMiner*, un software que apoyará en este proyecto para la extraer la información más exacta, algunas ventajas que presenta son: interfaz amigable, fácil entendimiento, aplicado a un negocio, esto con el fin de favorecer en la toma de decisiones en la agencia.

El ciclo de vida de un proyecto de minería de datos, es representado en la figura 3.1, donde define la metodología CRISP-DM, con las tareas a cumplir en cada etapa.

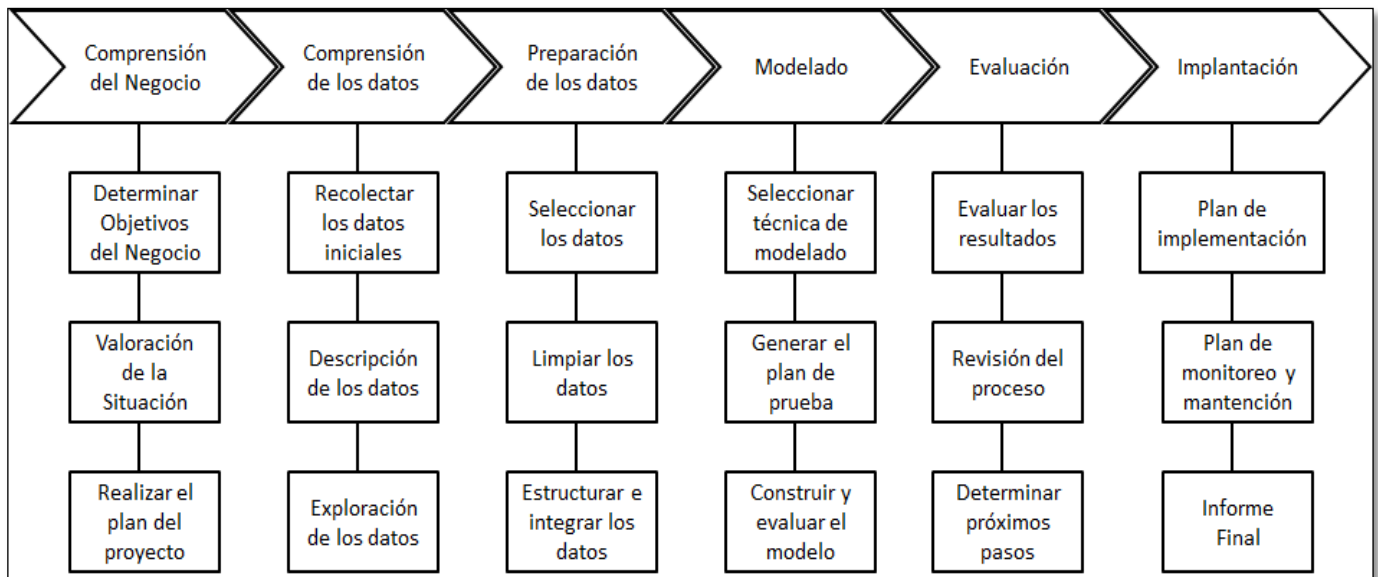


Figura 3.1. Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Aunque se muestran las etapas de forma lineal, este proceso puede ser aplicado varias veces hasta conseguir los resultados deseados, no es necesario que se realice de forma rígida, es decir, no limita a que una etapa tenga que estar precedida por la anterior, según el proyecto lo requiera.

3.1 Fases de la metodología CRISP-DM

3.1.1 Comprensión del negocio.

La fase uno se enfoca en entender los objetivos del proyecto y requerimientos desde una perspectiva del negocio, luego convertir este negocio en la definición de un problema de minería de datos y el diseño de un plan preliminar para conseguir el objetivo deseado, como se muestra en la figura 3.2.

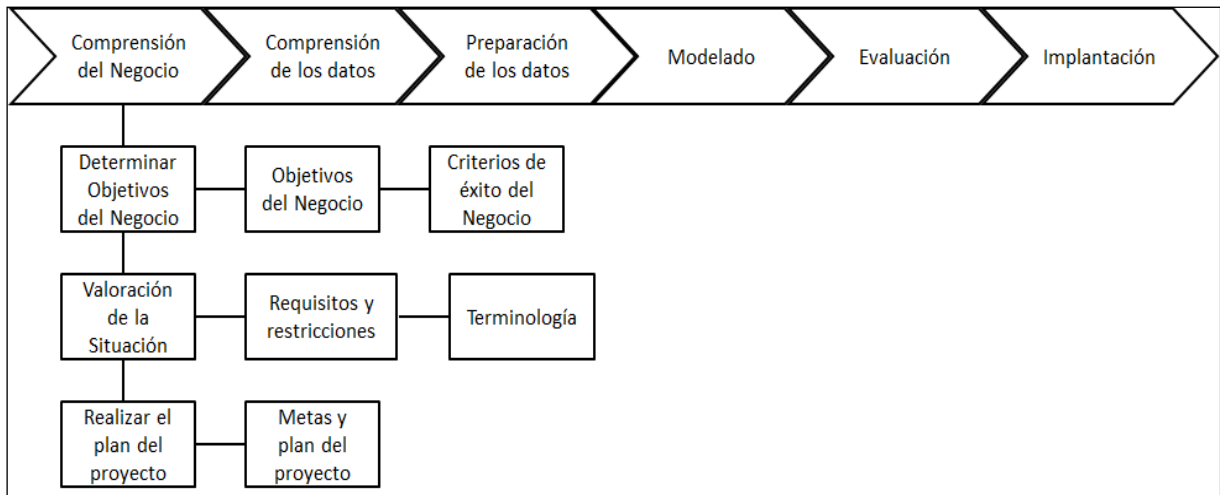


Figura 3.2. Fase 1 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Para esta fase se estructuró una entrevista, cuyo objetivo fue recabar información preliminar que nos permitió conocer las problemáticas que se presentan, así como la situación actual de este departamento, para definir los requisitos del problema y representar los objetivos del negocio en metas.

3.1.2 Fase de comprensión de los datos

La segunda fase comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis, como se muestra en la Figura 3.3, esta fase en conjunto de las dos siguientes, son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos.

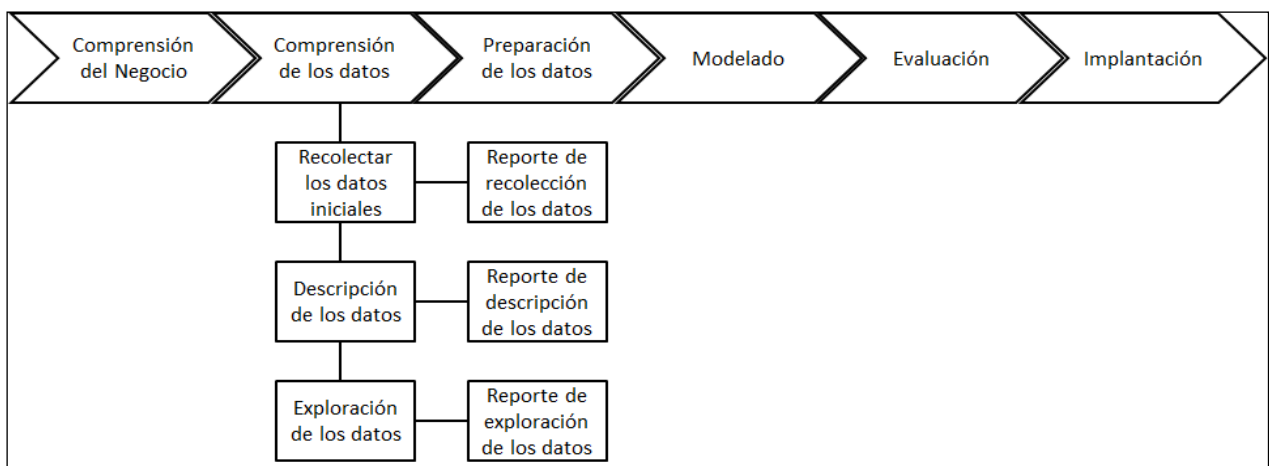


Figura 3.3. Fase 2 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Las principales tareas que se desarrollaron en esta base fueron: la recolección de los datos iniciales para ver su adecuación y procesarlos en un futuro, elaboración de un listado detallado para conocer los datos donde se describió su tipo, tamaño, entre otras, se tenía que definir un estándar entre los datos, también se obtuvo una estructura general de los datos, en apoyo con la base de datos que tiene la agencia para la obtención de información.

3.1.3 Fase de preparación de los datos

La fase de preparación de datos como se muestra en la figura 3.4, incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato, está relacionada con la fase de modelado.

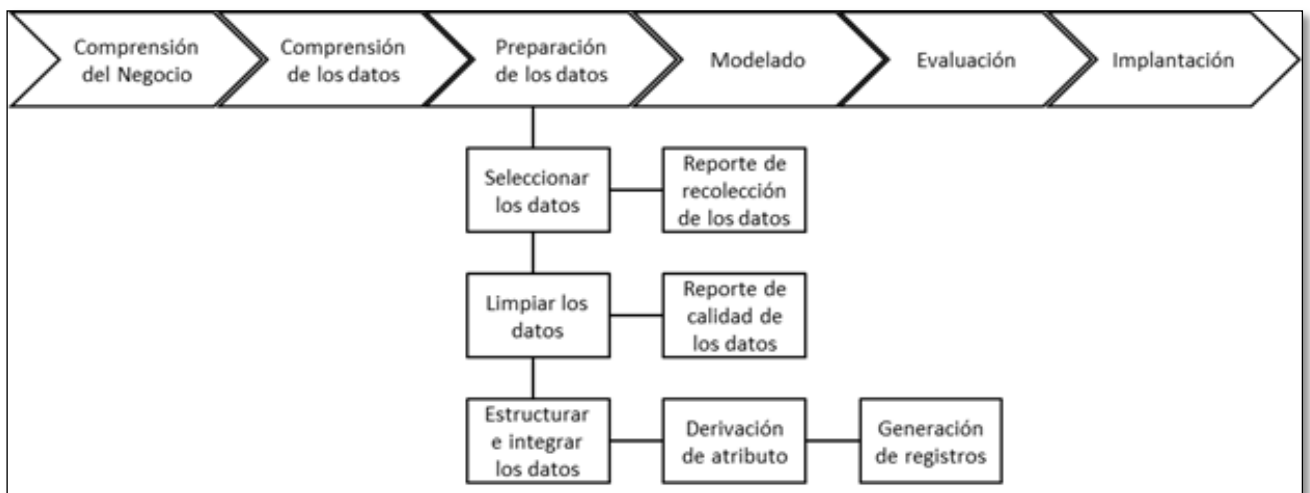


Figura 3.4. Fase 3 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Las tareas que involucraron esta fase fue obtener una selección de los datos los cuales fueron adquiridos de la fase anterior, nos apoyamos de criterios previamente establecidos, se tomó en cuenta la generación de nuevos atributos partiendo de los ya existentes en nuestra base de datos, así mismo se definió una nueva estructura en la base de los datos seleccionados, donde se fusionaron tablas y nuevos campos, ajustando los valores en los campos para mantener solo aquellos necesarios, los valores que no se requirieron se eliminaron.

3.1.4 Fase de modelado

En la fase cuatro de modelado y muestra la figura 3.5, se seleccionan las técnicas de modelado más apropiadas para el proyecto específico de Data Mining. Las técnicas a utilizar se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

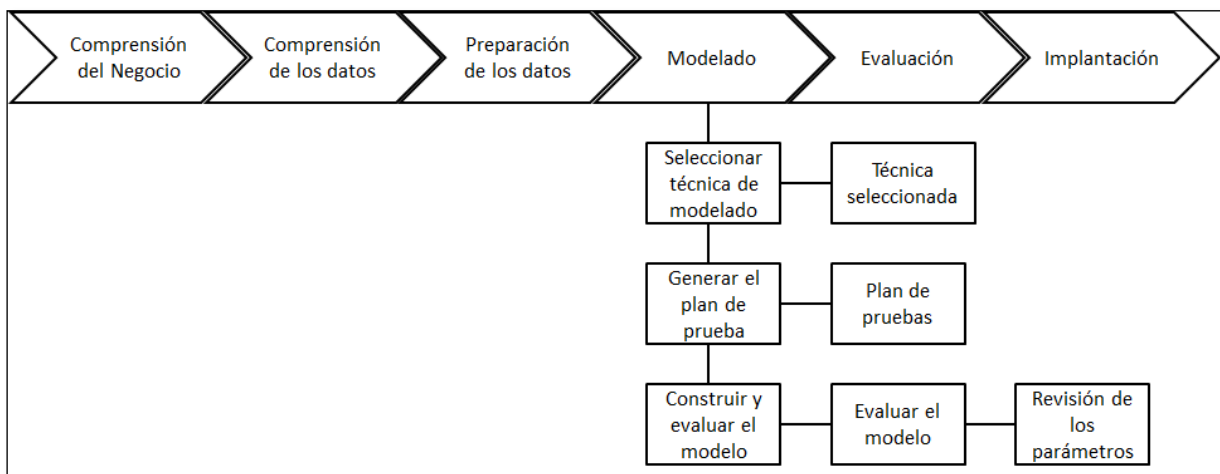


Figura 3.5. Fase 4 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Las actividad principal fue la selección de las técnicas a utilizar, así como la herramienta *RapidMiner* donde se crearán los modelos para obtención de resultados estos modelos crearán en base al objetivo principal del proyecto, se deben tener los datos ya preparados para generar uno o más modelos en apoyo con la herramienta para su interpretación, estos deben ser entendibles al usuario.

3.1.5 Fase de evaluación

En esta fase se evalúa el modelo, tener en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerar que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis como se muestra en la figura 3.6.

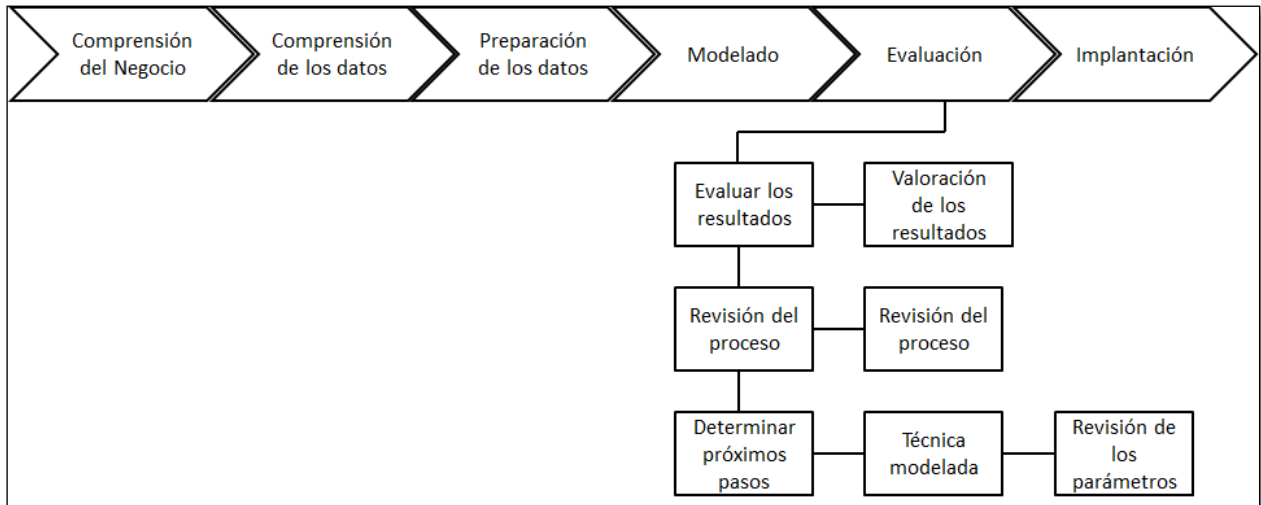


Figura 3.6. Fase 5 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

En esta etapa es preciso revisar si el modelo o los modelos construidos anteriormente, cumplan con los criterios para resolver la problemática en la que se presenta la agencia, se deberá verificar si los modelos hasta esta etapa han generado resultados satisfactorios, de ser que los modelos no cumplen con los establecidos se deberá repetir alguna etapa de las anteriores.

3.1.6 Fase de implementación

En la fase de implementación, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, al aplicar el modelo a diferentes conjuntos de datos o como parte del proceso como muestra la figura 3.7 por ejemplo, en análisis de riesgo crediticio, detección de fraudes, entre otras.

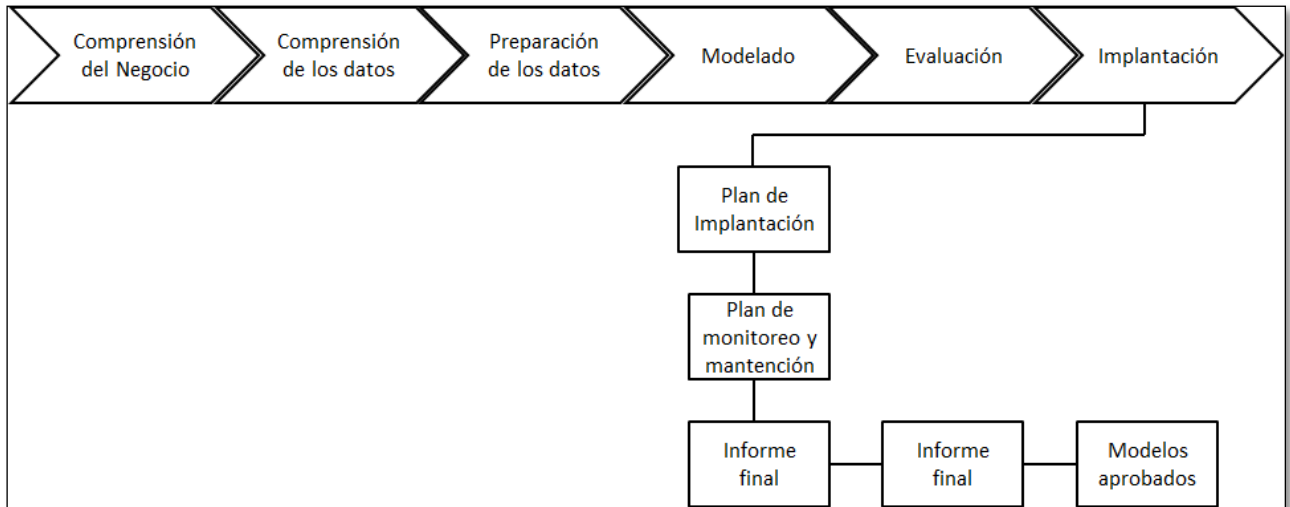


Figura 3.7. Fase 6 del Ciclo de vida de la metodología CRISP-DM (Fuente Propia)

Las actividades ejecutadas en esta fase fueron tomar el resultado de la evaluación para obtener resultados que beneficiarían en la toma de decisiones. Al tener el proceso de minería de datos ya implementada se deberá preparar la estrategia de monitoreo y mantenimiento para verificar que el modelo creado este arrojando información correcta.

3.1.6.1 Informe Final.

Como conclusión del proyecto de Minería de Datos, se desarrollará un informe con los puntos importantes del proyecto, donde se describirá los resultados logrados con la información de la agencia al indicar qué fue lo correcto e incorrecto, qué se realizó y posibles mejoras a futuro para su actualización.

CAPÍTULO IV.

DESARROLLO

4.1 Comprensión del negocio.

Se requiere identificar el comportamiento de las ventas en las sucursales, esto es de utilidad para la toma de decisiones para determinar o definir las ventas y reducción de estas, para esto se aplicó un cuestionario al departamento de ventas, para conocer la situación actual.

Los objetivos se pueden resumir en lo siguiente:

- Identificar unidades más demandadas en la agencia dependiente el tipo de unidad.
- Conocer qué tipo de cliente adquiere más unidades, es de importancia poder identificar si es persona física o moral.
- Determinar periodo donde las ventas tienen un aumento considerable.
- Verificar cuál de las tres empresas pertenecientes ha mantenido un alto número de ventas.

Contexto actual

En referencia a la situación global de la agencia se puede decir que toda la información necesaria para la explotación de datos se encuentra disponible en un sistema ERP (Enterprise Resource Planning) de forma dispersa, cada empresa maneja el mismo sistema de forma centralizada, el servidor se ubica en Tijuana. Cada sesión que manejan es distinta entre todas.

Al momento no se tiene en el medio un estudio en profundidad sobre las ventas en cada empresa existente.

Departamentos involucrados

El departamento de donde se tomará la información necesaria y también estará destinado el resultado es el de ventas, en el cual se concentra la mayor cantidad de información de donde se puede obtener el resultado deseado.

Área del problema

En el área de ventas no se tiene información detallada de las ventas, es requerido para conocer la situación actual en comparación con periodos anteriores, así mismo, tener estrategias que beneficien al departamento, siendo esta una desventaja el no contar con bases suficientes al tomar una decisión.

La minería de datos, debe ser forma para de utilización ya que no existen registros documentados de los comportamientos en las ventas, todo esto relacionado con los clientes y unidades demandadas.

Solución actual

Actualmente se recurre a consultas que indican muestras, por ejemplo SQL (Lenguaje de Consulta Estructurado) para obtener los datos necesarios, esta información la exportan a un documento Excel para realización de filtros para poder identificar cuantas ventas llevan al día por vendedor, solo conociendo la información del día

Objetivo primario de negocio

El objetivo principal en la agencia es contar con información completa y veraz para que el departamento se enriquezca y que pueda utilizar a su favor, ejempló el modelo más demandado entes los clientes.

Actividades

Cumplir con los objetivos propuestos por la aplicación de la metodología de explotación de datos dará respuesta a las siguientes preguntas:

- ¿Cuál es la unidad más demandada?
- ¿Qué modelos son los más solicitados?
- ¿Qué tipo de clientes son los que más adquieren unidades?
- ¿Qué empresa es la que tiene un alto nivel en las ventas?
- Importancia de tomar decisiones
- ¿En qué periodo se ha detectado un mayor aumento en las ventas?

4.1.1 Evaluación de la situación

Situación

Se cuenta con información detallada de las operaciones en el área, se tienen las bases de datos disponibles y funcionando en el sistema ERP (Enterprise Resource Planning), están siendo utilizadas y alimentadas diariamente.

Recursos de software

La agencia Automotriz cuenta con los siguientes sistemas en su operación diaria:

- Manejador de base de datos SQL, donde contiene la información al día del funcionamiento de la agencia, permitiendo exportar la información para su uso.
- Herramienta de Excel para realización de filtros que le ayudan a tener un panorama sencillo de sus ventas al día.
- Ejecutable de PDF para visualizar la información a un nivel gerencial.

4.1.2 Objetivos de minería de datos

Los objetivos en términos de minería de datos son:

- Definir el comportamiento de las ventas, a través de los modelos existentes y su demanda.
- Obtener un patrón del tipo de cliente con mejor adquisición.
- Obtener un resultado favorable donde el departamento obtenga un conocimiento que le favorecerá en la toma de decisiones.

Para que se cumplan los objetivos definidos, es necesario se realice los comparativos en periodos atrás que comprenden de año 2010 a la fecha.

Clasificación: Por año 2009, 2010, 2011, 2012.

Agrupamiento: Por año 2009, 2010, 2011, 2012.

Predicción: Por año 2009, 2010, 2011, 2012.

4.2 Comprensión y Análisis de Datos

4.2.1 Recolección de datos iniciales

Identificación de autos

Cada unidad automotriz es identificada con un número de artículo, el cual se especifica en cada uno de los registros que componen el presente trabajo.

Empresas

Cada empresa responsable es extraída con un código numérico.

Observación

Tiene como objetivo describir las ventas, este será un dato alfanumérico de N caracteres.

Fechas

Las fechas son extraídas en formato numérico con el formato aaaammdd: cuatro caracteres numéricos correspondientes al año, dos al mes y dos al día.

4.2.1.1 Requerimientos de información

La información necesaria para el logro de los objetivos del proyecto son los descritos en el punto anterior y lo mismos se encuentran disponibles en las base de datos.

Criterio de selección de los datos

Los atributos específicos que serán necesarios para la minería de datos son los que se muestran a continuación en la tabla 4.1:

Atributos
no_articulo
serie_unidad
nombre_unidad
des_unidad
Ano
Color
fecha_facturacion
no_vendedor
nombre_vendedor
no_cliente
nombre_cliente
tipo_cliente
Precio
Costo
Empresa
Almacén

Tabla 4.1. Muestra atributos a utilizar

Los datos se extraerán por periodos establecidos para realizar las particiones de datos extraídos de forma anual desde el año 2010 hasta la fecha actual.

Todos los datos son extraídos del manejador de bases de datos SQL Server, la estructura de la base de datos está normalizada, se puede apreciar Figura 4.1 el modelo entidad-relación de donde ser extraída la información.

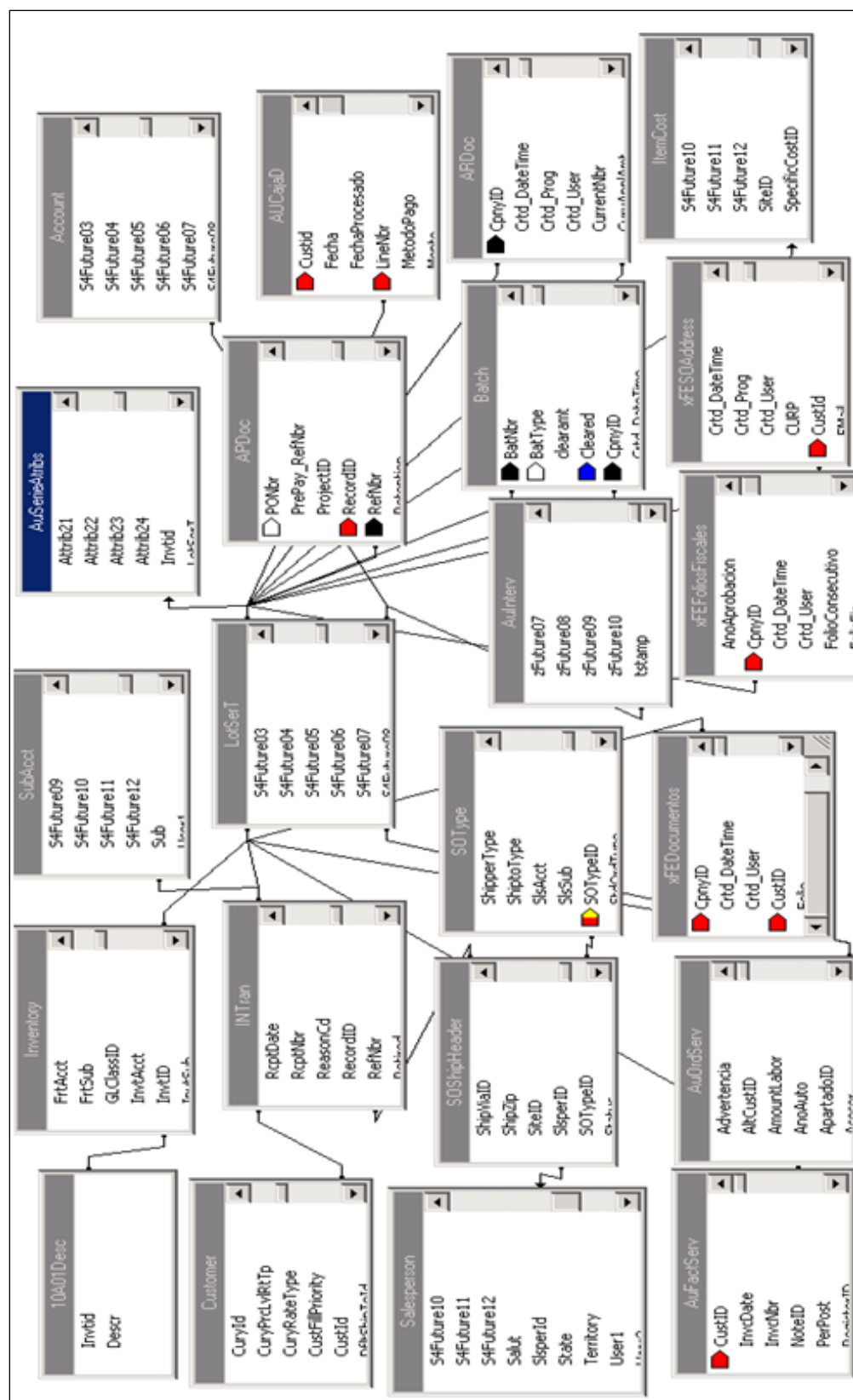


Figura 4.1. Entidad-Relación de estructura en SQL Server
(Fuente Propia)

Insertión y transformación de datos

Los datos que contienen sólo descripciones compuestas por caracteres alfanuméricos se procederán a relacionar.

4.2.1.2. Descripción de datos

En lineamientos generales los datos extraídos, que se muestran en la siguiente tabla, se han estandarizado en tipo de dato que permita su manipulación al momento de realizar la creación de modelos que serán utilizados para arrojar los resultados en este proyecto.

Campo	Tipo	Campo	Tipo
no_articulo	alfanumerico	Rfc	Alfanumérico
serie_unidad	alfanumerico	dirección	Texto
nombre_unidad	Texto	cantidad	Numérico
des_unidad	Texto	estado	Texto
Ano	numérico	Ciudad	Texto
Color	Texto	Aduana	Numérico
fecha_facturacion	Fecha	Motor	Alfanumérico
no_vendedor	numérico	Costo	Numérico
nombre_vendedor	Texto	empresa	Numérico
no_cliente	numérico	almacén	Numérico
nombre_cliente	Texto	Color	Texto
tipo_cliente	Texto	Modelo	Texto
Precio	numérico	Precio unidad	Alfanumérico

Tabla 4.2, muestra campo con su tipo de dato (Fuente Propia)

Datos y métodos de captura

Los datos son extraídos del manejador de base de datos SQL, para insertarlas en una hoja de cálculo Excel, en donde se trabaja a través de filtros para obtener información de fácil entendimiento

Si es necesaria la extracción de datos para continuar con su transformación, se cuenta con accesibilidad de los datos, a través de sentencias en SQL se podrá obtener información más detallada.

- Soshipheader-auserieatribs, la relación existente es la serie de la unidad para obtener la descripción completa de la unidad como serie, color, pedimento, entre otros.
- Soshipheader-itemcost, a través de la serie de unidad, se realiza la relación para mandar llamar los precios, costos.
- Soshipheader-company, con el id de la empresa se realiza la relación para identificar en que agencia procedió la venta.
- Soshipheader-lotsert, relación entre el número de factura de tabla de shoshipheader para conocer serial
- Shoshipheader-inventory, a través de id de almacén se conocer almacén afectado por compañía, esto ayudara a identificar si es un demo, seminuevo o nuevo.
- Soshipheader-soshippayments, con el id de pago, se identificara la forma de pago sea crédito o contado.

4.2.1.3 Exploración de datos

El algoritmo de construcción de árboles de decisión trabaja sobre la cantidad de registros tomados sin especificación de cantidad máxima y el objetivo inicial es el descubrimiento de patrones para el comportamiento de las ventas en las empresas, tipo de unidad, modelos, entre otras

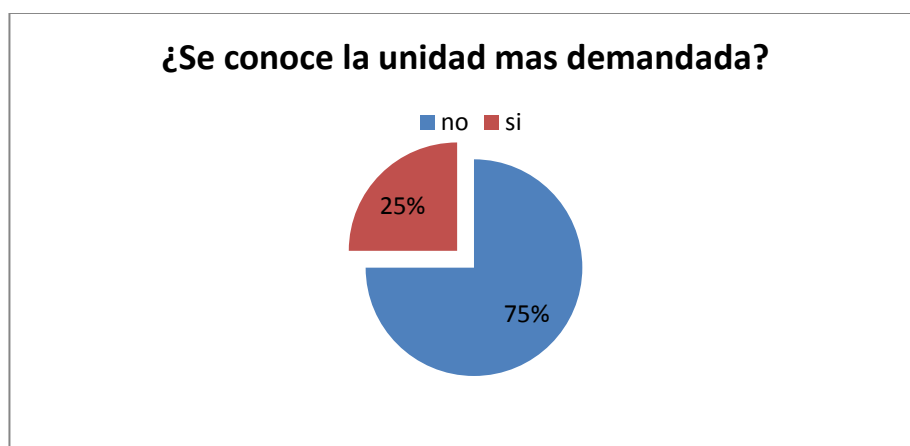
Informe de exploración de datos

Se realiza la exploración de datos sobre información extraída desde el año 2009 al 2012. Los primeros resultados van a mostrar información estadística y principalmente sirve para determinar la consistencia y completitud de datos.

También en esta etapa se podrá ver la necesidad de extraer de la base de datos atributos que posiblemente fueron omitidos y los cuales a partir de esta primera exploración se pueden vislumbrar.

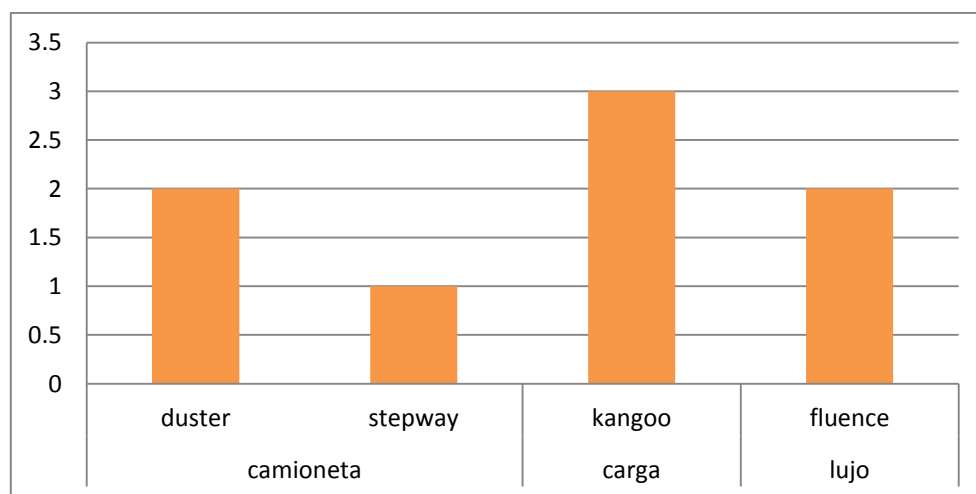
A continuación se muestra en las siguientes gráficas, el resultante de algunas preguntas de ejemplo que arrojó el cuestionario aplicado sobre la totalidad de la información con la que se desarrolla el presente proyecto, en base a la fase de exploración de datos inicial.

En la gráfica 1, se puede observar que el departamento involucrado desea saber que unidad es la más demandada para poder así tomar estrategias con las demás unidades.



Gráfica 1. Desconocimiento de unidades más vendidas
Fuente: Elaboración Propia

Así mismo, se observa en la gráfica 2, un comparativo de los diferentes tipos de unidades existentes según el modelo de unidad, donde usuario indica que unidad más solicitada es la Kangoo.



Gráfica 2. Tipo unidad según modelo, Fuente: Elaboración Propia

4.3 Preparación de Datos

4.3.1 Datos seleccionados

Los datos seleccionados para el análisis se muestran en la tabla 4.3, estos datos deben aun pasar por el proceso de limpieza y transformación.

Atributos		
campo	tipo	Descripción
cpnyid	numérico	identificador empresa
descr	carácter	descripcion unidad
invtid	carácter	no_articulo
lotsernbr	carácter	serie unidad
emailaddr	carácter	correo cliente
telephone	carácter	telefono cliente
attrib01	carácter	atributos unidad
attrib02	carácter	atributos unidad
attrib03	carácter	atributos unidad
attrib04	carácter	atributos unidad
attrib05	carácter	atributos unidad
attrib06	carácter	atributos unidad
slsperid	onumérico	identificador vendedor
custid	numérico	identificador cliente
taxid00	numérico	Impuestos
curydate	carácter	fecha venta
qty	numérico	Cantidad
addr1	carácter	dirección 1 cliente
addr2	carácter	dirección 1 cliente
anoauto	numérico	año unidad
costunit	numérico	costo unitario

Tabla 4.3. Datos seleccionados para el análisis. (Fuente Propia)

El volumen de datos a explotar varía en función del software a utilizar. En este proyecto se utilizará la herramienta *RapidMiner* para la creación de los modelos, la información a utilizar se extraerá de SQL Server, para la limpieza de los datos y transformación de los mismos se utiliza Excel de Microsoft.

4.3.2 Limpieza de datos

La base de datos con la se cuenta para el presente trabajo contiene información relacionada con operaciones de diferentes modelos de unidades, pero sólo algunas manejan movimientos, por este motivo, se decidió eliminar los modelos sin información, aquellos datos que fueron eliminados pueden ser objeto de un estudio posterior para continuar con esta línea de investigación en la industria automotriz.

En la figura 4.4 se muestra la generación de la primera estructura de datos resultado de la limpieza de datos.

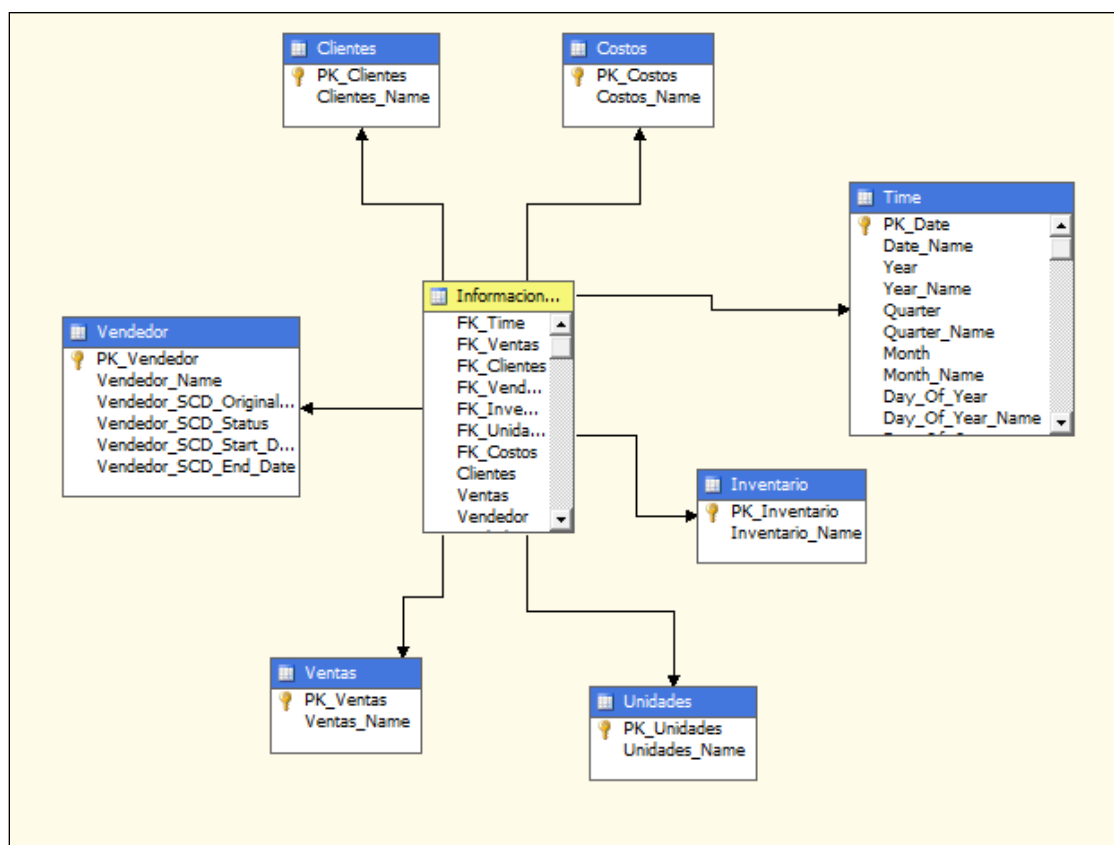


Figura 4.4. Estructura 1 resultante de la limpieza de datos. (Fuente Propia)

Por motivos de normalización para el ingreso de información a los modelos de software que se tienen se utiliza herramienta SQL para su codificación, puesto que los mismos están disponibles en la base de datos en formato de caracteres alfanuméricos.

4.3.3 Construcción de datos

Como se mencionó en el punto anterior se ha realizado la transformación de los atributos para utilizarse en los modelos a realizar, al definir un tipo de dato estándar, debido a que se encontraban con un tipo de dato no estandarizado, complicando su uso para la creación de relaciones de datos más sencillas.

Lo recomendable para el proceso de transformación es la programación de una consulta SQL con las sentencias correspondientes para la transformación de los datos, todos los datos extraídos previos a la incorporación de los mismos en el software de análisis de datos pasan a un archivo de planilla de cálculo.

La explotación de los datos se realizará sobre información codificada en diferentes formatos de datos, luego de realizar el estudio del significado de los resultados se procede a la captura del significado en la base de datos. Los números de serie es el identificador que contiene cada unidad que permite realizar un filtro apropiado puesto que es único y universal para cada una de las mismas. La estructura del número de identificación del auto permite saber dentro de sí mismo a que marca y modelo pertenece, por lo tanto en cada consulta SQL se incorpora la sentencia correspondiente para beneficiar el filtro.

4.3.4 Integración de datos

Los datos extraídos se combinan a partir de tablas relacionadas por sus respectivas claves, dando como resultado una nueva estructura de datos que muestra la creación de nuevos atributos significativos a utilizar, esta estructura apoyara en la creación de modelos, en la figura 4.5 muestra lo mencionado.

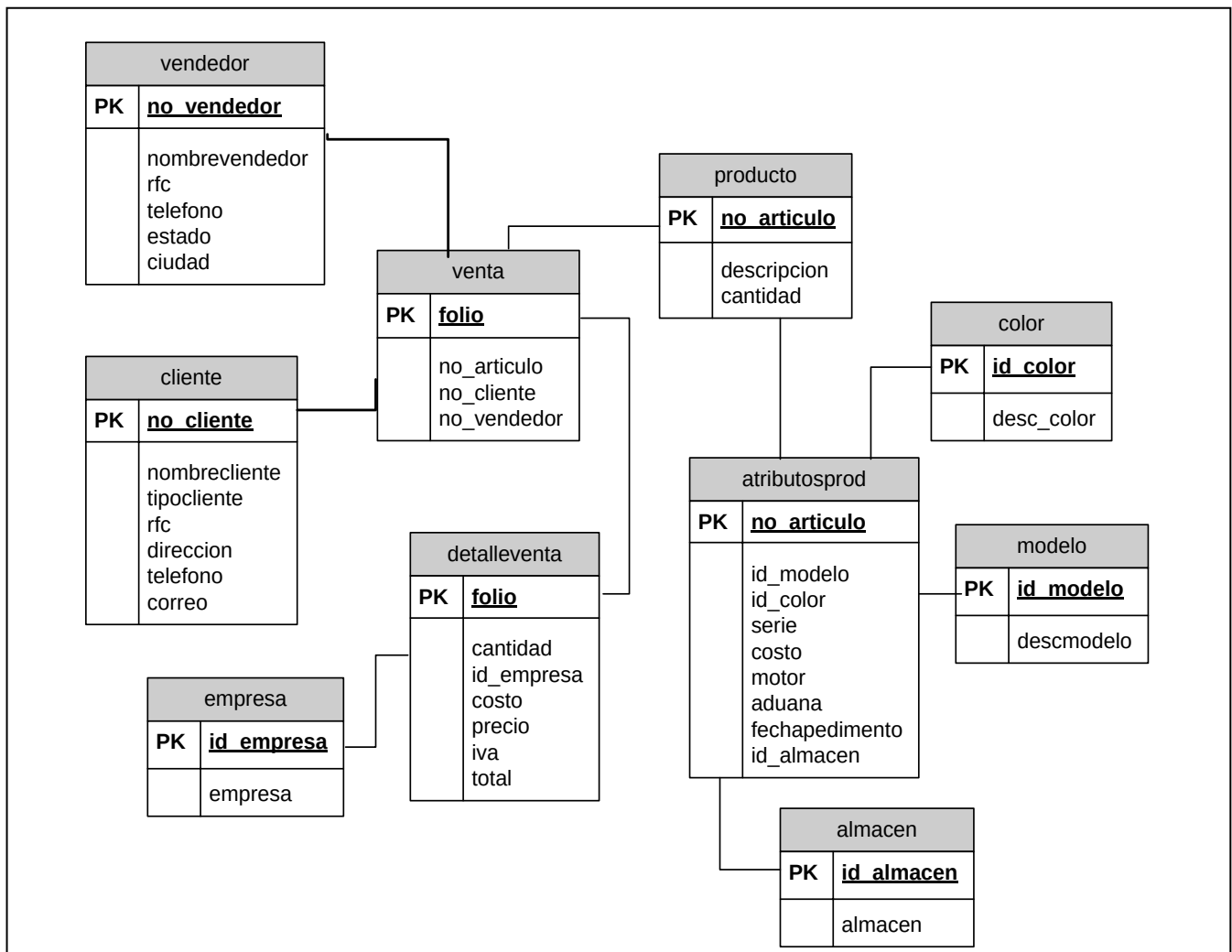


Figura 4.5. Estructura de datos final. (Fuente Propia)

Los atributos finales se obtienen de las siguientes tablas:

- Vendedor.
- Cliente.
- Empresa.
- Venta
- Detalleventa.
- Producto.
- Atributosprod
- Almacén.
- Modelo.
- Color.

4.3.5 Formateo de datos

En cuanto al orden de las tablas, no se requiere una posición específica para los efectos de normalizar, a continuación se define un orden del conjunto de tablas:

Tabla Empresa			
campo	Tipo	tamaño	Descripción
id_empresa	numérico	5	identificador empresa
empresa	Carácter	15	nombre empresa

Tabla 4.4. Tabla empresa con sus atributos. (Fuente Propia)

Tabla Detalleventa			
campo	Tipo	tamaño	Descripción
folio	numérico	5	identificador venta
cantidad	numérico	5	cantidad surtir
id_empresa	numérico	5	identificador empresa
costo	numérico	10	costo unitario
precio	numérico	10	precio de venta
IVA	numérico	10	impuesto aplicado
total	numérico	10	total pagar

Tabla 4.5. Tabla detalle venta con sus atributos. (Fuente Propia)

Tabla Producto			
campo	Tipo	tamaño	Descripción
no_articulo	numérico	15	identificador articulo
descripcion	Carácter	25	descripcion de articulo
cantidad	numérico	5	cantidad existente

Tabla 4.6. Tabla producto con sus atributos. (Fuente Propia)

Tabla Atributosprod			
campo	tipo	tamaño	Descripción
no_articulo	numérico	15	identificador articulo
id_modelo	numérico	5	identificador modelo
id_color	numérico	5	identificador color
serie	carácter	17	serie unidad
costo	numérico	10	costo unitario
motor	carácter	5	clave motor
aduanas	numérico	10	clave aduana
fechapedimento	carácter	10	fecha pedimento
id_almacen	numérico	5	identificador almacén

Tabla 4.7. Tabla atributosprod con sus atributos. (Fuente Propia)

Tabla Color			
campo	tipo	tamaño	Descripción
id_color	numérico	5	identificador color
desc_color	carácter	15	descripcion color

Tabla 4.8. Tabla color con sus atributos. (Fuente Propia)

Tabla Modelo			
campo	tipo	tamaño	Descripción
id_modelo	numérico	5	identificador modelo
descmodelo	carácter	15	descripcion modelo

Tabla 4.9. Tabla modelo con sus atributos. (Fuente Propia)

Tabla Almacén			
campo	tipo	tamaño	descripcion
id_almacen	numérico	5	identificador almacén
almacén	numérico	5	numero almacén

Tabla 4.10. Tabla almacén con sus atributos. (Fuente Propia)

Tabla Vendedor			
campo	tipo	tamaño	descripcion
no_vendedor	numérico	5	identificador vendedor
nombrevendedor	carácter	25	nombre de vendedor
rfc	carácter	13	rfc vendedor
telefono	carácter	10	telefono de vendedor
estado	carácter	15	estado donde radica
ciudad	carácter	10	ciudad donde radica

Tabla 4.11. Tabla vendedor con sus atributos. (Fuente Propia)

Tabla Cliente			
campo	tipo	tamaño	Descripción
no_cliente	numérico	5	identificador cliente
nombrecliente	carácter	25	nombre de cliente
tipocliente	carácter	13	tipo de cliente (moral o física)
rfc	carácter	13	rfc de cliente
dirección	carácter	30	dirección de cliente
correo	carácter	20	correo electrónico
telefono	carácter	10	telefono cliente

Tabla 4.12. Tabla cliente con sus atributos. (Fuente Propia)

Tabla Venta			
campo	tipo	tamaño	Descripción
folio	numérico	5	identificador venta
no_articulo	carácter	15	identificador articulo
no_cliente	numérico	5	identificador cliente
no_vendedor	numérico	5	identificador vendedor

Tabla 4.13. Tabla venta con sus atributos. (Fuente Propia)

4.4 Modelado

Para obtener un resultado hacia las ventas, es necesario contar en la herramienta Rapid Miner para crear los modelos que arroja la información necesaria para apoyo en las estrategias a tomar. La figura 4.6 muestra la estructura de la creación del primer modelo.

El modelo que se presenta fue utilizado para definir un estándar de los datos dentro de la herramienta.

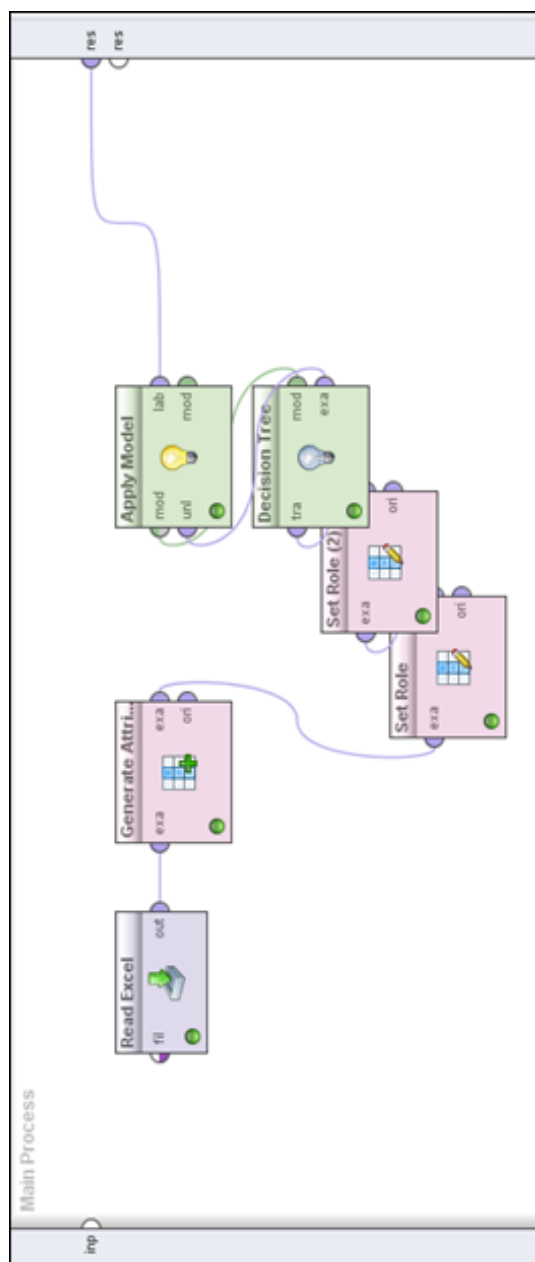


Figura 4.6. Primer modelo estandarización de datos
(Fuente Propia)

Toda la información presentada a continuación es la que se extrajo del ERP, se trabajó en su limpieza para su uso, la figura 4.7 presenta el resultante de todo el trabajo realizado en las primeras fases de la metodología.

☒ Data View
 ☐ Meta Data View
 ☐ Plot View
 ☐ Advanced Charts
 ☐ Annotations

ExampleSet (826 examples, 829 special attributes, 28 regular attributes)

View Filter (826 / 826): all

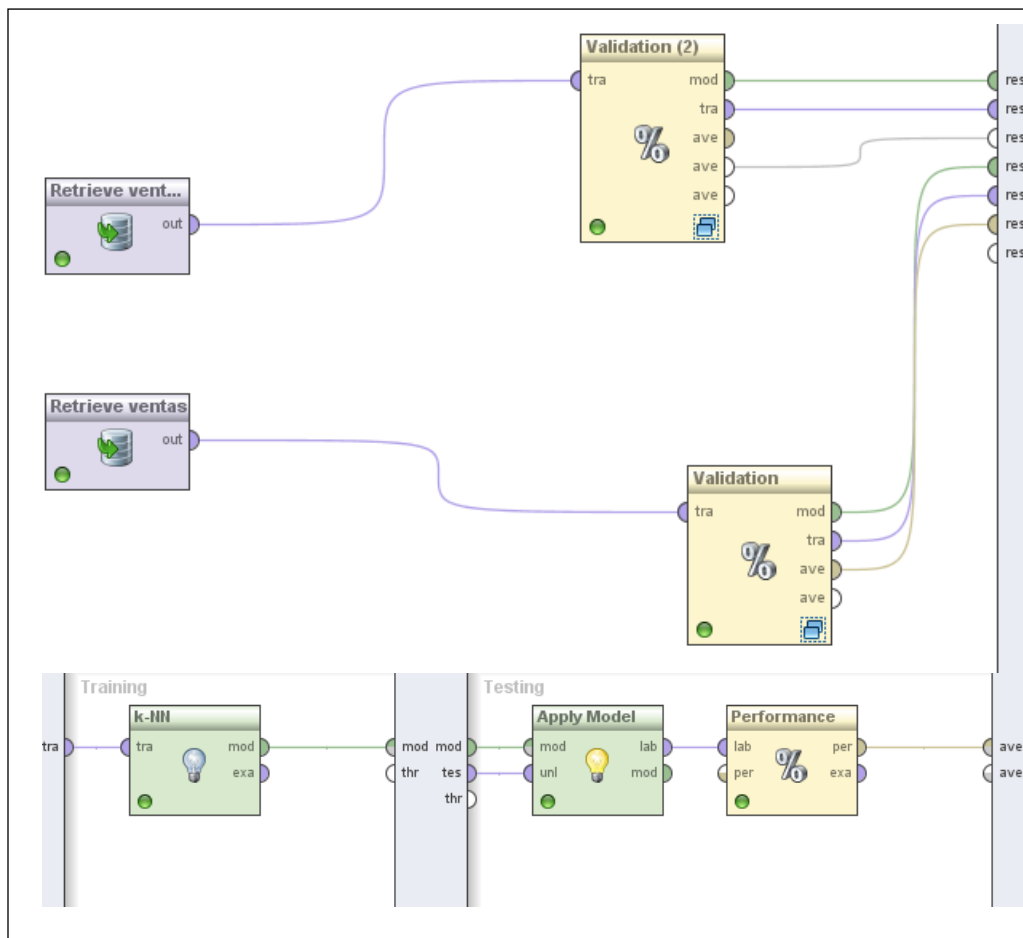
confidence(...prediction(e...	lote	direccion1	direccion2	ciudad	pais	nombre	telefono	estado	cp	compañía	nombrec...	fecha
?	MA0000696	8038	VIANA 6401	VILLA RESIC ?	MX	ROSAL	6642525687	BC	22664	10	TIJUANA	30/01/2013
0	MA0000696	13400	ACACIAS 11	LOS PINOS	MX	MURIEL GAF	6865683510	BC	21230	20	MEXICALI	12/06/2013
0	MA0000603	7037	AVE MARIA	EL LIENZO I	MX	ALEJO DUR	6865645046	BC	21258	20	MEXICALI	28/12/2012
0	MA0000708	14159	DE LA JOLLA	VILLA COLO	MX	LOPEZ SOTI	6861656232	BC	21298	20	MEXICALI	29/06/2013
0	MA0000637	8764	RIO PECOS	HACIENDA I	MX	GARCIA AGF	6868412692	BC	21300	20	MEXICALI	27/02/2013
0	A000057	1771	DR COSSIO	CHAPULTEPEC	MX	RODRIGUEZ	6646815416	BC	22045	10	TIJUANA	28/06/2012
0	MA0000710	14173	PUENTE LA	VILLAS RES	MX	SEGURA GL	6862728130	BC	21600	20	MEXICALI	29/06/2013
0	A000054	1709	SAN JUAN C	RESIDENC	MX	CORAZA SE	6641651102	BC	22194	10	TIJUANA	26/06/2012
0	MA0000474	895	IZAPA 3700	MAYAKHAN I	MX	VALLE LOPE	6865651300	BC	21376	20	MEXICALI	31/05/2012
0	MA0000501	2151	CALLE F 32	NUEVA	MX	KING TONEL	6865519292	BC	21100	20	MEXICALI	12/07/2012
0	MA0000463	587	RIO SANTA I	BUGAMBILJA	MX	PEREZ GON	6865625584	BC	21395	20	MEXICALI	23/05/2012
0	MA0000458	459	ZILIA 883	MONTECAR	MX	TORRES VA	6865683606	BC	21255	20	MEXICALI	18/05/2012
0	MA0000468	649	ALMERIA 96	CONJUNTO	MX	MURILLO RI	6865578135	BC	21350	20	MEXICALI	28/05/2012
0	MA0000505	2397	BOSQUE DE	BOSQUES I	MX	AVILA RAMIF	6865611830	BC	21720	20	MEXICALI	20/07/2012
0	MA0000519	2626	LORETO 66	CATAVINA	MX	ESPINOZA J	6869056163	BC	21240	20	MEXICALI	31/07/2012
0	MA0000446	21	GUADALUP	PROHOGAR	MX	BARAJAS PIL	6865646882	BC	21240	20	MEXICALI	3/05/2012
0	MA0000516	2558	AVE. RIO CC	VILLA VERD	MX	VERDUGO S	6865623440	BC	21395	20	MEXICALI	27/07/2012
0	MA0000611	7789	AVE. JALAPA	LIBERTAD	MX	GOMEZ SAL	6535346992	SR	83487	20	MEXICALI	16/01/2013

Figura 4.7. Estructura con los campos a utilizar en Rapid Miner
(Fuente Propia)

4.4.1 Selección de las técnicas de modelado

En el modelado se trabajará con herramienta Rapid Miner donde se espera definir diferentes técnicas para clasificación de información que beneficiaran al departamento, a continuación se muestran las técnicas utilizadas en la investigación.

Una de ellas es la **Técnica de K-NN**, proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos y que fueron extraídos de la base de datos, por tal motivo se debe crear modelo correspondiente como se muestra en la figura 4.8.



4.8 Modelo de técnica K-nn. (Fuente Propia)

En la figura 4.9 se observa la clasificación existente de las agencias, las cuales serían el patrón para poder identificar por empresa sus ventas.

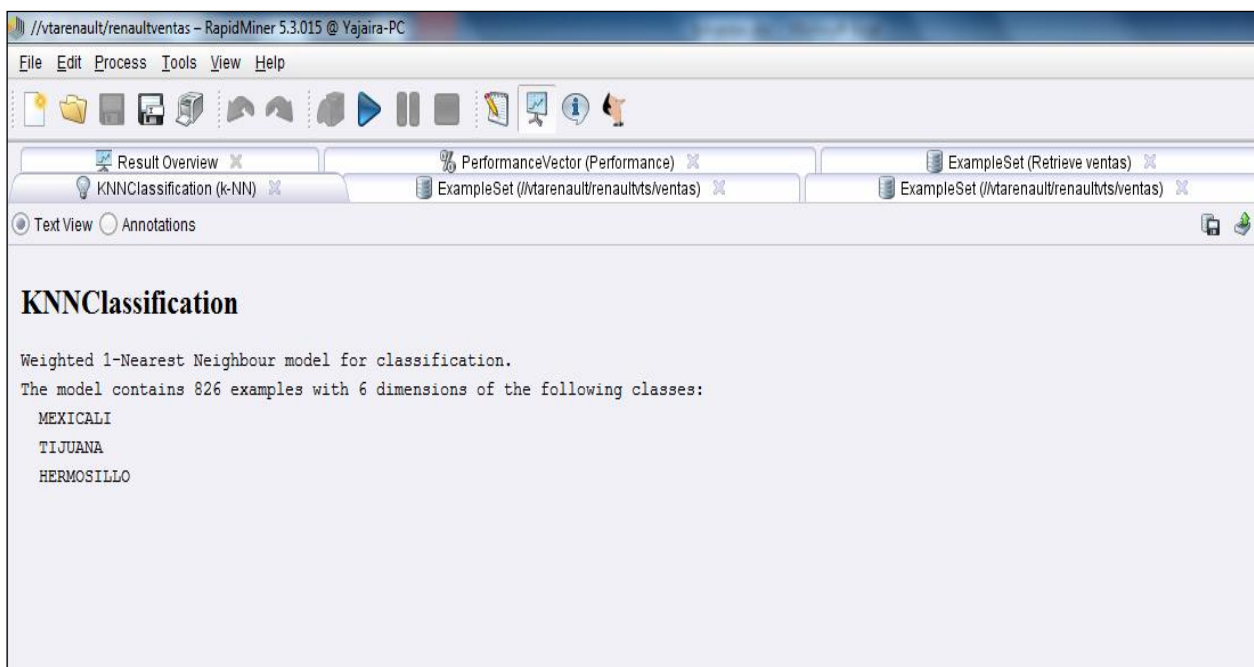


Figura 4.9 Clasificación KNN por agencia. (Fuente Propia)

La figura 4.10, muestra el resultado de la técnica aplicada por empresas indicando las ventas reales con porcentaje de precisión.

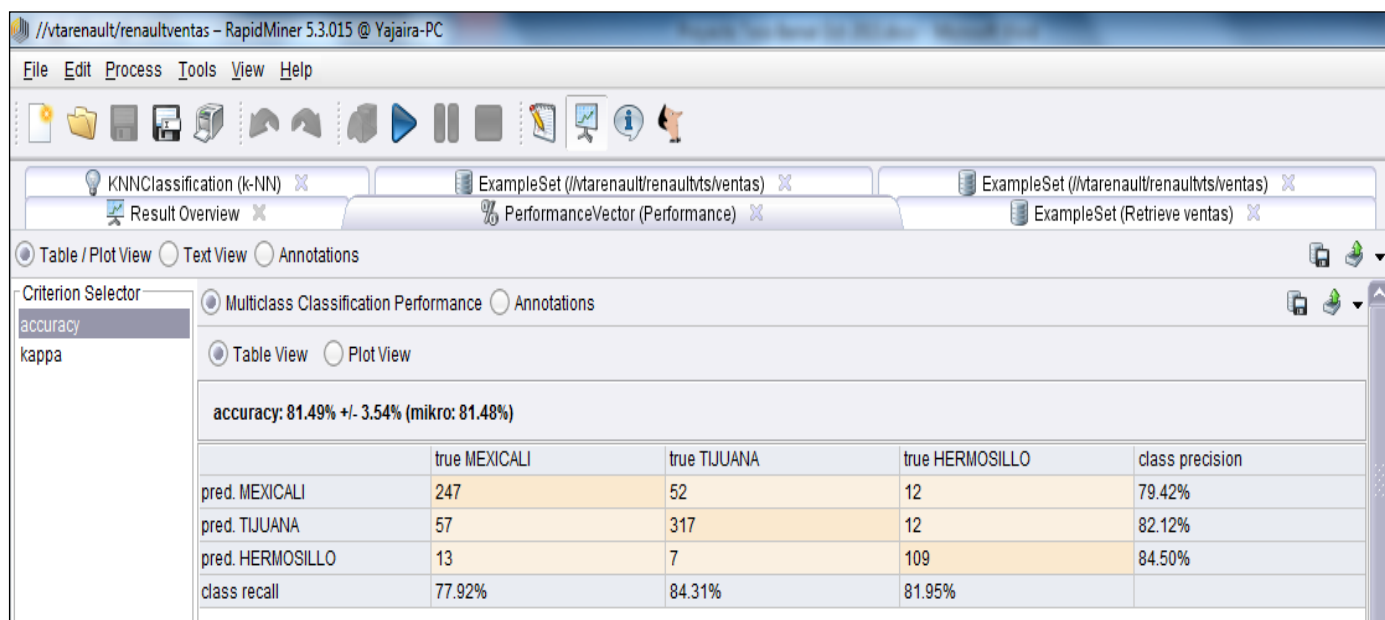
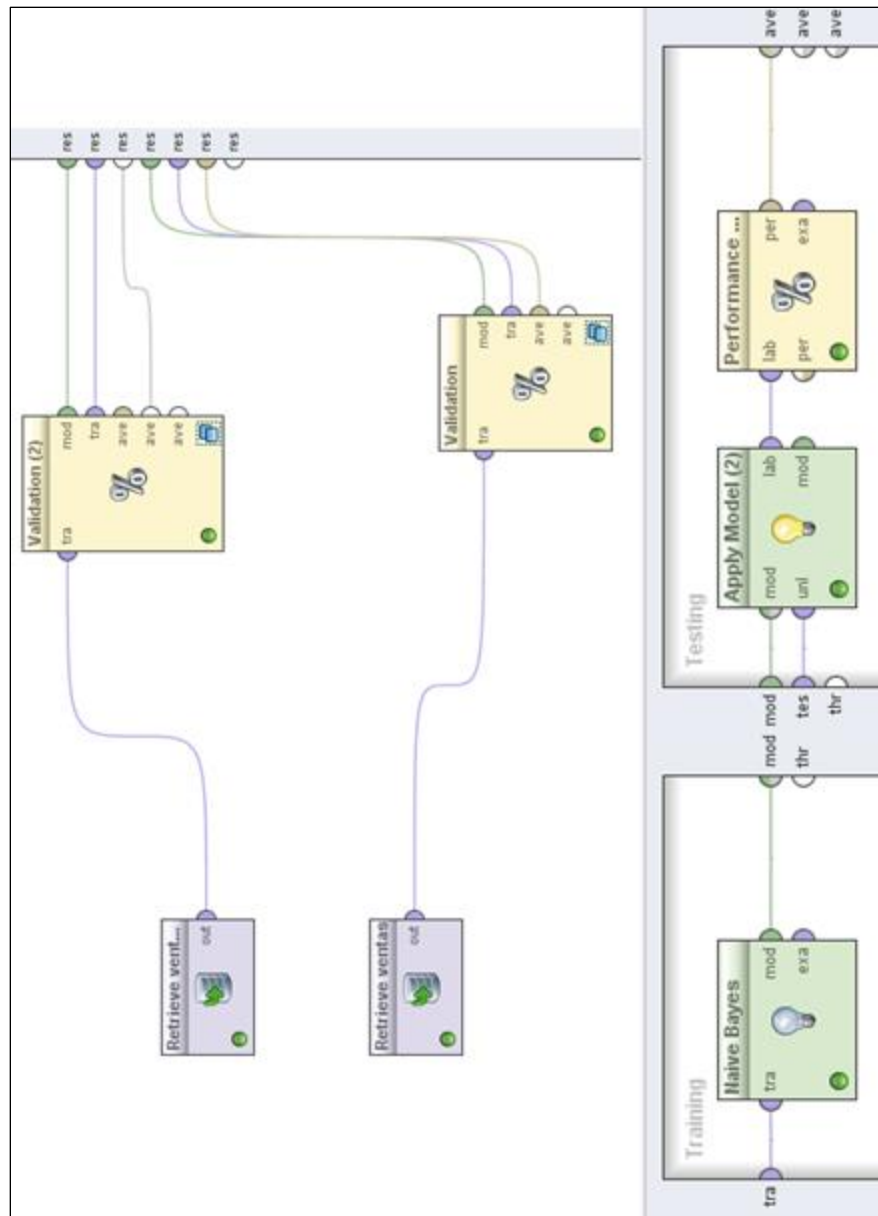


Figura 4.10. Porcentaje de precisión (Fuente Propia)

Técnica de Bayes es un proceso que permite abordar en forma diferente el área de la toma de decisiones, la información se debe basar en cuanto a la ocurrencia de un evento, a continuación en la figura 4.11 se visualiza la creación del modelo través de la herramienta de Rapid Miner.



4.11 Modelo de técnica Bayes
(Fuente Propia)

Una distribución simple por la clase de agencia, es representada en la siguiente figura 4.12.

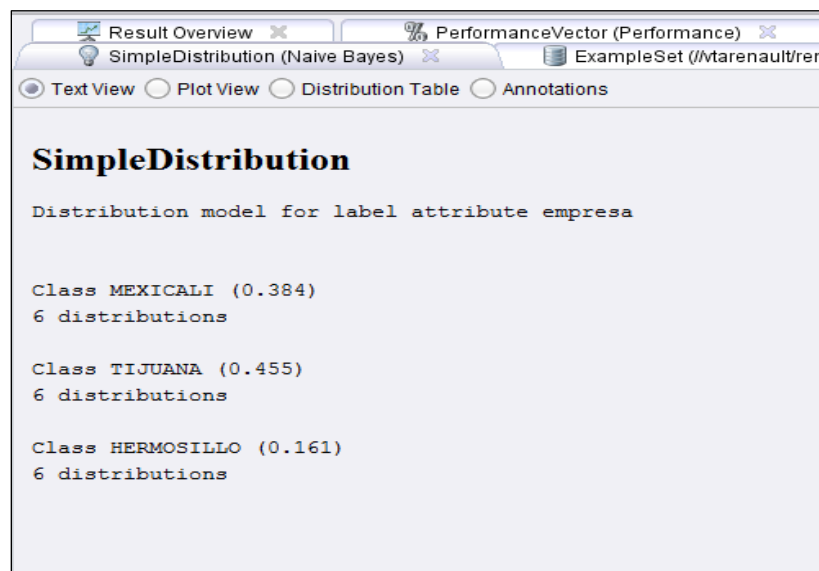


Figura 4.12. Clasificación Bayes por agencia. (Fuente Propia)

Modelo de Clustering

El análisis es realizado agrupando: Unidad-Venta-Lugar-Fecha, Unidad-Venta-Modelo-Lugar, Unidad-Venta-Modelo-Cliente.

El agrupamiento de partes, ventas, tipo de unidad y clientes permite visualizar el comportamiento de las ventas separadas por periodos, así mismo determinando el tipo de cliente y el tipo de unidad, este análisis se realiza por periodos, seguido del lugar de la venta, que tipo de unidad fue la más vendida.

El proceso de agrupamiento es realizado con matrices cuadradas de 10 x 10 en todos los casos. A continuación se muestra resumen de los resultados obtenidos a partir la ejecución del primer período.

Unidad-Cantidad-Lugar (Período 1 – Año 2011)

	Unidad	version unidad	lugar	cantidad
Clúster 1	Duster	Dynamique	Tijuana	0
Clúster 2	Duster	Expression	Tijuana	0
Clúster 3	Kangoo	Carga	Tijuana	350
Clúster 4	Fluence	Dynamique	Tijuana	277
Clúster 5	Fluence	Expression	Tijuana	245
Clúster 6	Safrane	Dynamique	Tijuana	354

Clúster 7	Safrane	Expression	Tijuana	234
Clúster 8	Koleos	Dynamique	Tijuana	234
Clúster 9	Koleos	Expression	Tijuana	123
Clúster 10	Stepway	Dynamique	Tijuana	145
Clúster 11	Stepway	Expression	Tijuana	235

Unidad-Cantidad-Lugar (Período 1 – Año 2011)

	Unidad	version unidad	lugar	cantidad
Clúster 1	Duster	Dynamique	Mexicali	0
Clúster 2	Duster	Expression	Mexicali	0
Clúster 3	Kangoo	Carga	Mexicali	379
Clúster 4	Fluence	Dynamique	Mexicali	234
Clúster 5	Fluence	Expression	Mexicali	241
Clúster 6	Safrane	Dynamique	Mexicali	134
Clúster 7	Safrane	Expression	Mexicali	134
Clúster 8	Koleos	Dynamique	Mexicali	212
Clúster 9	Koleos	Expression	Mexicali	113
Clúster 10	Stepway	Dynamique	Mexicali	134
Clúster 11	Stepway	Expression	Mexicali	122

Tabla 4.14. Clúster por versión (Fuente Propia)

CAPÍTULO V.

RESULTADOS

Los resultados extraídos con la herramienta *RapidMiner*, son desglosados a continuación y estos se basan de las técnicas y modelos antes creados y aplicados.

Como primer resultado la herramienta arroja, con apoyo de la técnica K-nn, una estructura de Clúster por Empresa, al mostrar la actividad total de las ventas, pero entre más color tenga, indica que tiene más actividad como puede observar en la figura 5.1

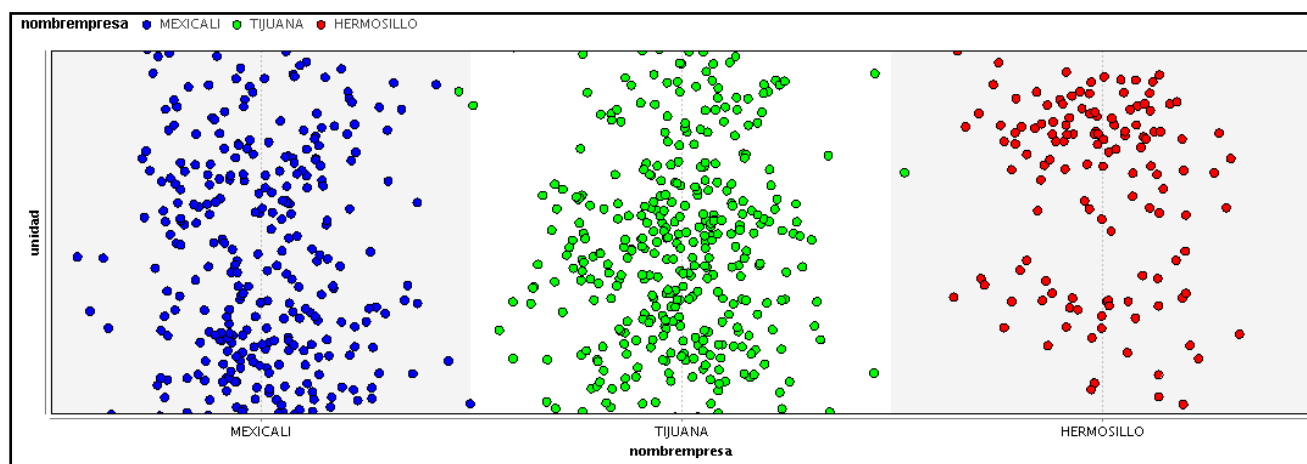


Figura 5.1. Muestra por empresa. (Fuente Propia)

Con la técnica K-nn se puede obtener una estructura de Clúster por una segmentación, la figura 5.2 muestra esta estructura en base a ventas por periodo, que comprende ventas por mes y por empresa, arrojando en que mes las ventas tuvieron mejor afluencia.

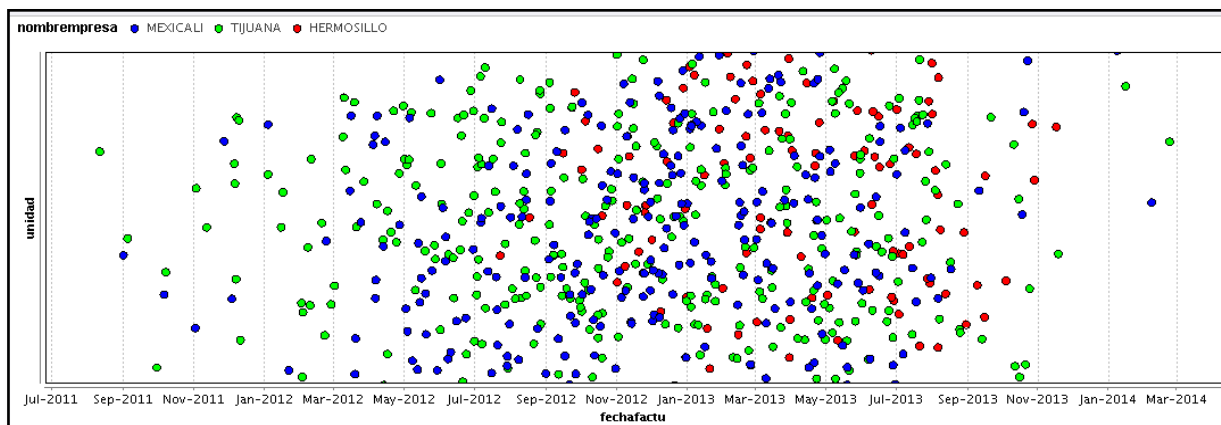


Figura 5.2 Muestra por periodo, empresa y unidad. (Fuente Propia)

Para tener un mejor entendimiento de la Figura 5.3, se muestra a continuación las versiones del año 2012 correspondiente a la venta por versión, se puede apreciar que la unidad más demandada de dicho periodo es: KANGOO EXPRESS AC 1.6 T/M 2012, este resultado es obtenido de la empresa de Tijuana.

Unidad	Cantidad
DUSTER DYNAMIQUE PACK 2.0 LT 2013	11
DUSTER DYNAMIQUE TA 2.0 LT 2013	21
DUSTER DYNAMIQUE TM 2.0 LT 2013	6
DUSTER EXPRESSION TA 2.0 LT 2013	16
DUSTER EXPRESSION TM 2.0 LT 2013	14
FLUENCE AUTHENTIQUE 2.0L TM 2012	16
FLUENCE DYNAMIQUE 2.0L CVT 2012	3
FLUENCE DYNAMIQUE PACK 2.0L CVT 2012	6
FLUENCE EXPRESSION 2.0L CVT 2012	14
FLUENCE PRIVILEGE SERIE LIMITADA 2012	3
KANGOO EXPRESS 1.6 T/M 2012	8
KANGOO EXPRESS AC 1.6 T/M 2012	45
KOLEOS DYNAMIQUE OPENSKY T/A CVT 2012	1
KOLEOS DYNAMIQUE T/A CVT 2012	1
KOLEOS EXPRESSION T/A CVT 2012	5
SANDERO 1.6 DYNAMIQUE TM 2013	3
SANDERO 1.6 EXPRESSION TM 2013	7
SANDERO 5 PUERTAS 1.6 DYNAMIQUE T/A 2012	6
SANDERO 5 PUERTAS 1.6 EXPRESSION T/M 2012	1
SCALA 4 PUERTAS 1.6 DYNAMIQUE T/A 2012	1
SCALA 4 PUERTAS 1.6 DYNAMIQUE T/A 2013	2
SCALA 4 PUERTAS 1.6 DYNAMIQUE T/M 2013	1
SCALA 4 PUERTAS 1.6 EXPRESSION T/A 2013	5
SCALA 4 PUERTAS 1.6 EXPRESSION T/M 2013	3
SCALA 4 PUERTAS 1.6 EXPRESSION T/M 2012	4
Seminuevos Otras Marcas Personas Fisicas	1
Seminuevos Renault Personas Fisicas	2
STEPWAY 1.6 TM 2013	9
STEPWAY 5 PUERTAS 1.6 DYNAMIQUE T/M 2012	11
TRAFIC VU L2H2 T/M 2012	9

Tabla 4.13. Representación periodo 2012 Tijuana, por versión. (Fuente Propia)

En la figura 5.4 Muestreo de ventas por unidad, se puede visualizar que en conjunto desglosa como resultado las unidades más demandadas en cada empresa, como resultado de las técnicas aplicadas del periodo 2013.

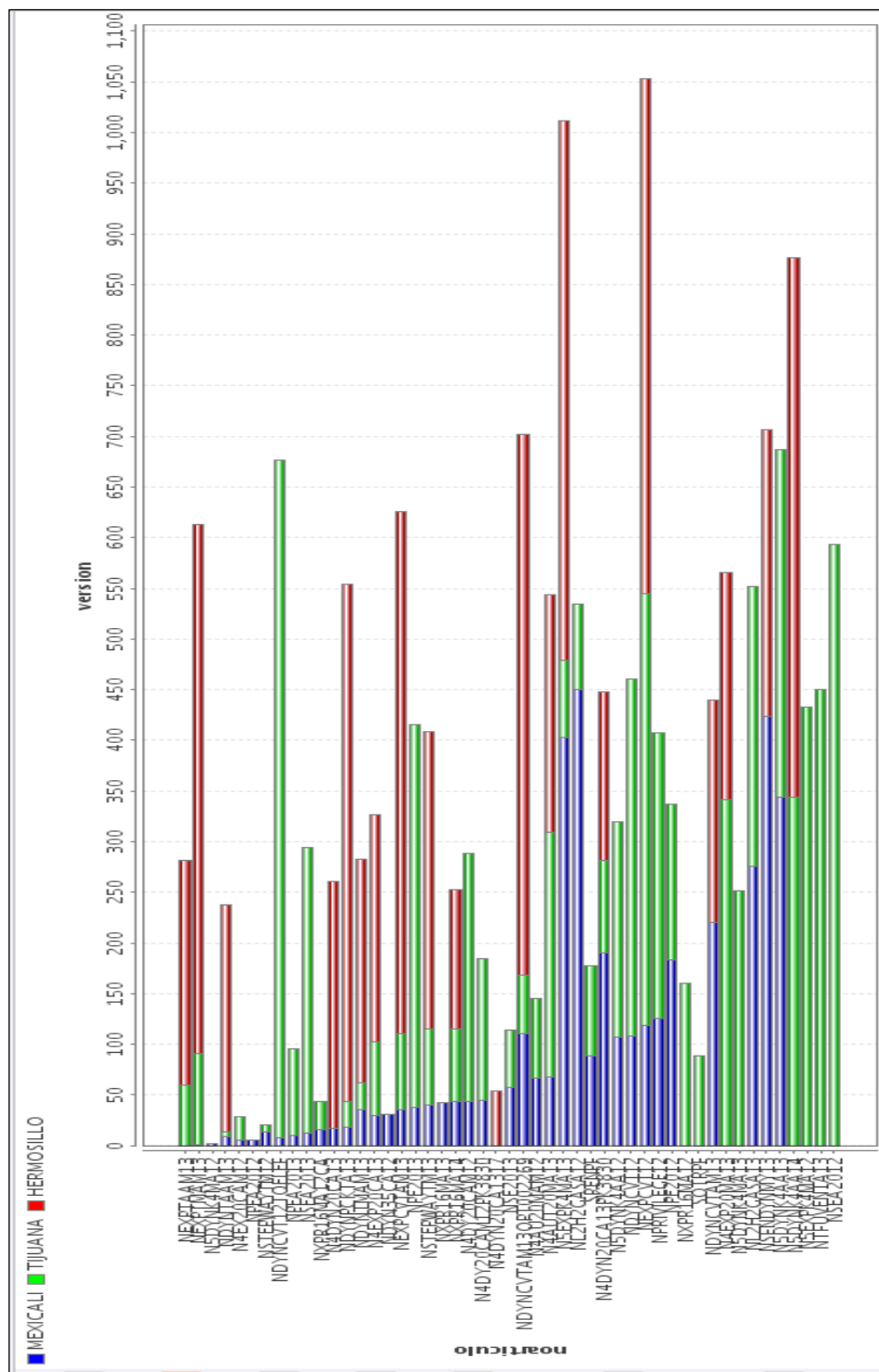


Figura 5.4 Muestro de ventas por unidad (Fuente Propia)

Se puede mostrar a continuación en la Tabla 4.16 cantidad más marcada de acuerdo a la Figura 5.4 del periodo 2013.

Unidad	Cantidad
DUSTER DYNAMIQUE PACK 2.0 LT 2013	13
DUSTER DYNAMIQUE TA 2.0 LT 2013	14
DUSTER DYNAMIQUE TM 2.0 LT 2013	8
DUSTER EXPRESSION TA 2.0 LT 2013	24
DUSTER EXPRESSION TM 2.0 LT 2013	12
FLUENCE AUTHENTIQUE 2.0L TM 2012	1
FLUENCE AUTHENTIQUE 2.0L TM 2013	6
FLUENCE DYNAMIQUE 2.0L CVT 2012	1
FLUENCE DYNAMIQUE PACK 2.0L CVT 2012	2
FLUENCE DYNAMIQUE PACK 2.0L CVT 2013	3
FLUENCE EXPRESSION 2.0L CVT 2012	1
FLUENCE EXPRESSION 2.0L CVT 2013	3
FLUENCE EXPRESSION 2.0L TM 2013	2
KANGOO EXPRESS AC 1.6 T/M 2014	30
KOLEOS DYNAMIQUE OPEN SKY 2013	2
KOLEOS DYNAMIQUE T/A CVT 2012	3
KOLEOS EXPRESSION 2013	1
SANDERO 1.6 DYNAMIQUE TA 2013	1
SANDERO 1.6 DYNAMIQUE TA 2014	3
SANDERO 1.6 DYNAMIQUE TM 2013	1
SANDERO 1.6 EXPRESSION TM 2013	3
SCALA 4 PUERTAS 1.6 DYNAMIQUE T/M 2013	1
Seminuevos Otras Marcas Personas Físicas	5
STEPWAY 1.6 TM 2013	35
TRAFIC VU L2H2 T/M 2012	3
TRAFIC VU L2H2 T/M 2013	1
TRAFIC VU L2H2 T/M GASOLINA AC 2013	2

Tabla 4.16. Representación periodo 2013 ventas por versión de unidad. (Fuente Propia)

Con la técnica Bayes se pueden visualizar la información de diferentes formas, para verificar cual ha sido el vendedor con mejores ventas se aplica esta técnicas mostrando el resultado en la figura 5.5, indicando en este caso, Ma. Guadalupe Vásquez sería mejor vendedor.

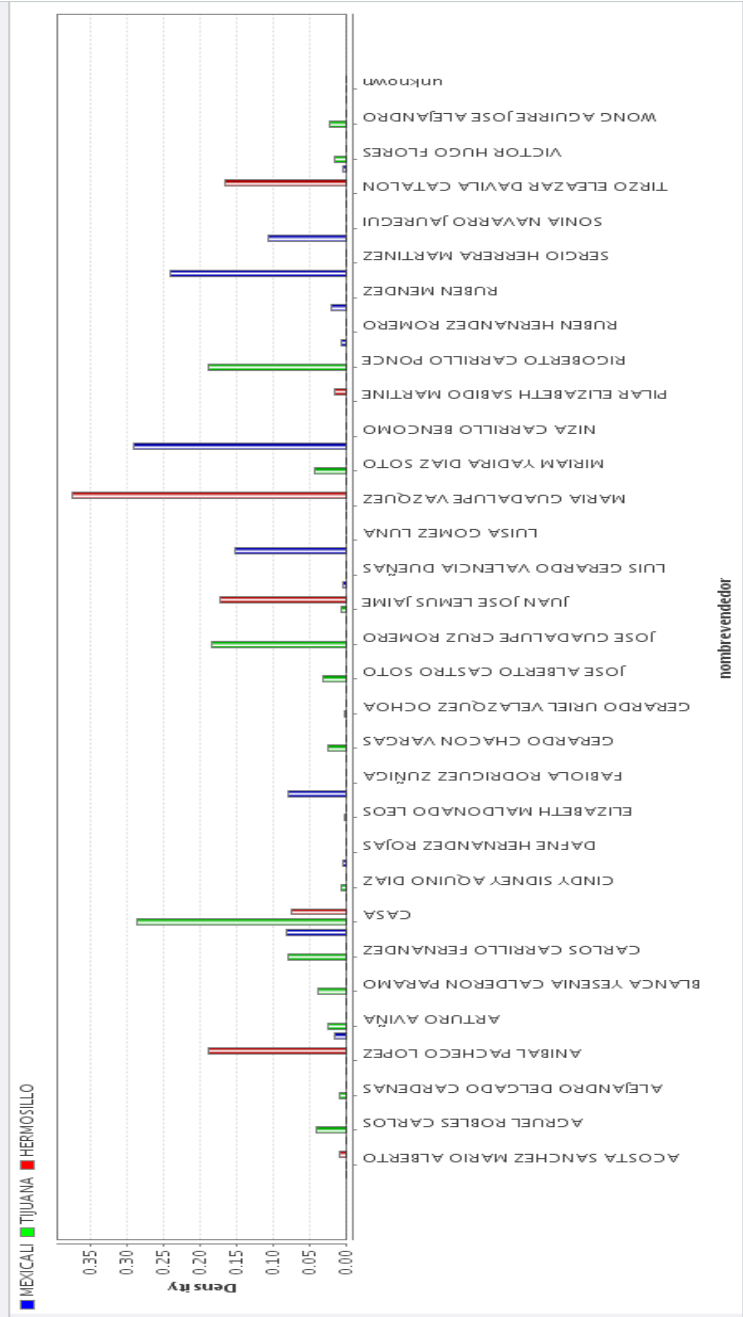


Figura 5.5. Muestra por periodo, empresa y unidad
(Fuente Propia)

Cada resultado se puede visualizar en un formato PDF para que nivel gerencial pueda presentar esta información como reportes, a continuación en la figura 5.6 se muestra la representación de este.

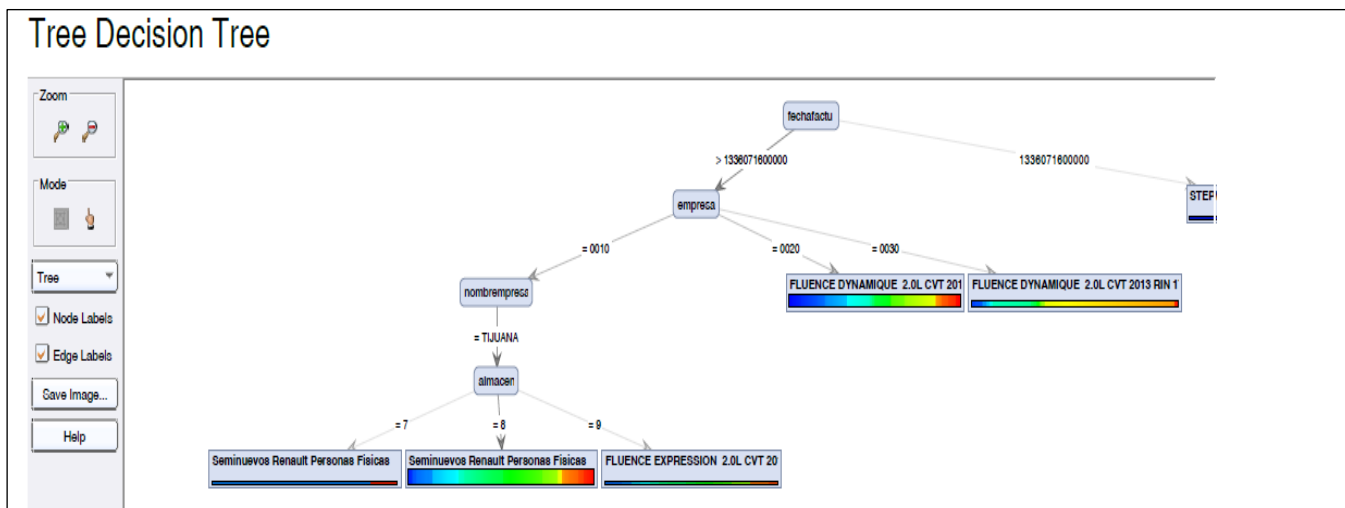


Figura 5.6. Decisión Árbol en presentación en PDF. (Fuente Propia)

Se pueden mostrar diferentes diseños, en diferentes formatos según la presentación que usuario desee mostrar.

La exportación se puede realizar desde la herramienta, donde usuario puede realizar la selección del formato que requiera como muestra la Figura 5.7.

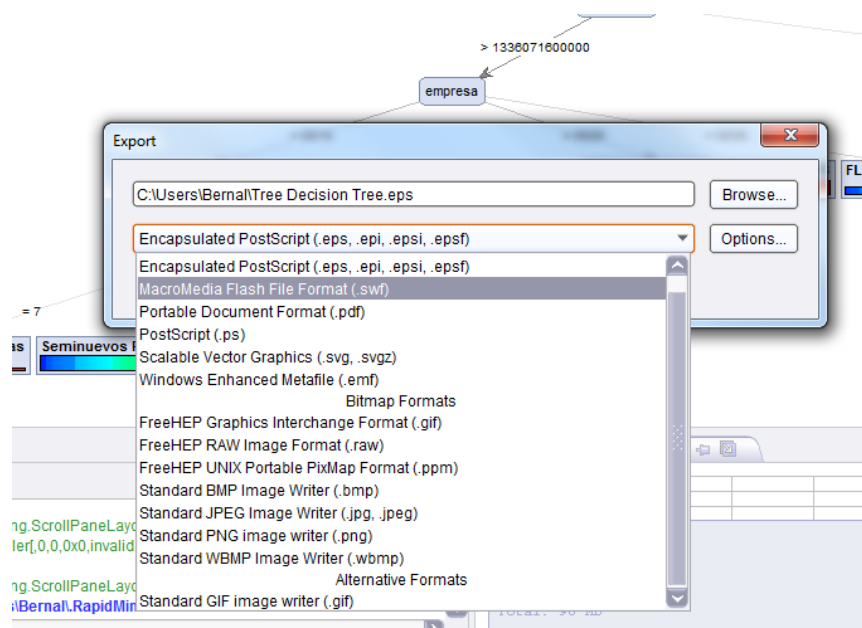


Figura 5.7. Selección de formato para extracción de reporte. (Fuente Propia)

CAPITULO VI.

CONCLUSIONES

La información es un elemento importante en toda organización en especial para aquellas compañías que manejan grandes volúmenes de información y el realizar consultas convencionales en las bases de datos no es suficiente para resolver problemas de negocios, una herramienta como la minería de datos que es un conjunto de herramientas y técnicas de análisis de datos permiten crear escenarios, de los cuales se puede obtener información útil para la toma de decisiones a nivel gerencial, a través de cada técnica implementada en la minería de datos consistió en identificar diferentes patrones.

Para el desarrollo de este proyecto se analizaron dos metodologías para la aplicación de la minería de datos, estas metodologías son SEMMA y CRISP-DM que pueden implementarse a través de sus distintas fases que se encuentran para producir un desarrollo iterativo e interactivo.

El problema detectado fue que no se contaba con información al momento que apoye en la toma de decisiones y estrategias que beneficien al departamento de ventas de la agencia automotriz Renault, siendo nuestro objetivo principal implementar una herramienta de minería de datos que permita obtener información de tiempos determinados a través de una metodología aplicada para la toma de decisiones.

A través de un análisis de las metodologías para el desarrollo de la minería de datos, la metodología CRISP-DM se consideró como la apropiada para comprensión, explotación de datos, atender la problemática planteada, por pensar desde sus primeras fases en conocer el negocio y favorecerlo.

Una de las ventajas por la cual es apropiado trabajar con la metodología mencionada es que es independiente a las herramientas de software, siendo una particularidad entre las otras metodologías existentes. El llevar cada fase como lo detalla la metodología esta se vuelve flexible y completa en su proceso, como base metodológica para futuros proyectos en la industria automotriz, por la existencia de diferentes estándares y sistemas de captura de información y la generalidad que presenta la metodología CRISP-DM le permite hacerse

recomendable para futuras implementaciones. Es importante llevar cada fase paso a paso, donde el proceso de la minería de datos genera conocimiento por medio de la depuración, enriquecimiento y transformación de datos para formar una estructura amigable y creación de modelos para su evaluación a través de software propuesto *RapidMiner*.

Si bien los resultados del trabajo muestran patrones de comportamiento satisfactorios es interesante ver que se pueden realizar más combinaciones para la explotación de datos en lo que se refiere a los lugares, modelos de autos o periodos de registros introducidos en cada modelo construido en el software.

Se puede decir que cualquiera que sea el problema a resolver no existe una única técnica para solucionarlo, sino que puede ser resuelto con distintas herramientas y técnicas, por ejemplo Excel, Weka (herramienta de minería de datos), entre otras.

La aplicación del Árbol de clasificación es totalmente óptima para este proyecto y en futuros proyectos de minería de datos de esta industria. Inclusive independientemente de la distribución o el agrupamiento de la información es de importancia utilizarlos al inicio de la explotación de datos para tener una referencia de la situación de las cantidades de los datos.

También se pueden aplicar la técnica de K-nn y Bayes, que permiten seguir, analizar todos los momentos del proyecto para verificar la coherencia con lo que se está trabajando, estas técnicas apoyan para construcción de diagramas que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

En cuanto a Clustering es necesario estudiar la posibilidad de trabajar con matrices de menor tamaño, debido que este estudio requiere mucho tiempo de computación para procesamiento de información, se aplicó esta técnica para poder verificar los periodos de ventas, ventas por tipo de empresa y unidad, es decir que la combinación de posibilidades en función a la información disponible es muy alta y cabe destacar que lo más importante de este proyecto es demostrar

la viabilidad de líneas de investigación futuras con la aplicación de técnicas de explotación de datos.

La razón principal por la que la agencia desea implementar una minería de datos es precisamente conocer a detalle su productividad en las ventas, esto a nivel gerencial, la agencia maneja un ERP que como todo sistema provee reportes, aun así dista de ser una solución tan completa como una minería de datos al trabajar con técnicas de precisión.

La herramienta *RapidMiner* es una plataforma para uso gerencial en apoyo a departamento de sistemas para la generación de reportes, siendo un complemento ideal para entregar soluciones que apoyen al tomar decisiones.

Al tener un panorama por periodos y datos más concretos, departamento de ventas se encuentra analizando los resultados para ver los beneficios de la información.

La implementación de la minería de datos ayuda a obtener resultados satisfactorios por medio de software y técnicas de clasificación, destacando que la agencia automotriz logro conocer por periodos las ventas sobresalientes, así como por unidad y su empresa perteneciente, aunque es necesario continuar trabajando en futuras implementaciones con la metodología CRISP-DM en el negocio, para mejorar la explotación de información. Con lo anterior se concluye que se logró el objetivo planteado en este proyecto.

Recomendaciones Futuras

Desarrollo de la metodología con información extraída de base de datos con otros estándares de inspección de unidades. Se recuerda que cada compañía puede tener su propio estándar de inspección el cual define el formato en el que los registros pueden ser explotados.

Con la aplicación del Modelo de Clasificación se puede extraer todo el conjunto de reglas de negocio generado para ser aplicado en el software tanto de captura de datos como en la aplicación que alimenta la base de datos. Esto es, a partir de las reglas generadas por el modelo validar todos los datos que ingresan al sistema.

Contar con una aplicación que extraiga la información de RapidMiner para uso del vendedor, o bien capacitar a personal para utilizar herramienta, está diseñada esta herramienta para generar reporte a nivel gerencial para tener bases y resultados para tomar decisiones y estrategias favoreciendo al departamento de ventas.

BIBLIOGRAFÍA

Aguilar, J. A. (7 de Octubre de 2011). Aplicacion de la mineria de datos para la toma de decisiones.

Ayuso, G. M. (1997). Importancia informatica de las bibliografias de bibliografias. *Interamericana de bibliotecnologia*.

Berry y Linoff. (2004). *Data Mining Techniques: for marketing, sales, and customer relationship management*. Indiana, EU: 2da edición por Wiley Publishing.

Booch, G. (2005). *El Lenguaje Unificado de Modelado*. Massachussets: Addison Wesley.

Castillo, G. G. (2008). *Desarrollo de un modelo basado en técnicas de Minería de Datos para clasificar zonas climatológicamente similares en el estado de Michoacan*. Michoacan.

Date, C. (2001). *Introducción a los Sistemas de Bases de Datos* (Septima Edición ed.). Pearson Educación.

Flores, I. H. (2009). *Deteccion de Patrones de daños y averias en la industria automotriz*. Buenos Aires.

Gallardo, J. A. (2003). *Metodologia para la definicion de requisitos en proyectos de Data Mining*. Mexico: EPB Sistemas de Conocimientos.

Galviz, M. (2004). *Confrontacion de dos tecnicas de mineria de datos aplicadas a un dominio especifico*. Bogota.

Gilberto Lorenzo Martinez Luna, Adolfo Guzman Arenas. (1994). *Busqueda de Patrones de Comportamiento en Cubos de Datos* (Vol. 1). Mexico: Software Pro International e IDASA.

Hand, Mannila y Smyth. (2001). *Principles of Data Mining*. Estados Unidos: The Massachusetts Institute of Technology Press.

Hernandez, F. B. (2005). Tesis Desarrollo de un Almacen de Datos para el Control Vehicular. Mexico D.F.

Hillermann, W. (2000). *Toma de Decisiones*. Guatemala.

Huamantumba, R. (Agosto de 2007). Manual para diseño y desarrollo de Datamart.

Intelesis. (2007). *Agencia Automotriz*. Recuperado el 5 de Diciembre de 2012, de <http://www.intelisis.com/erp/automotriz.html>

Kendall, K. E., & Kendall, J. E. (2005). *Analisis y Diseño de Sistemas* (Sexta Edición ed.). Mexico: Pearson Educacion.

Marcano, Aular Yelitza Josefina y Talavera, Pereira Rosalba. (2007). *Mineria de Datos como soporte a la toma de decisiones empresariales* (Vol. 23). Venezuela: Zulia.

Mari, G. (19 de Septiembre de 2009). *Taller Digital*. Universidad de Alicante. Recuperado el 10 de Diciembre de 2012, de <http://blog.eltallerdigital.com/2009/09/la-toma-de-decisiones-y-la-organizacion-del-trabajo-en-la-empresa-reflexiones-en-torno-a-ellas/>

María Moreno García, Luis Miguel Quintales, Francisco García Peñalvo y José Polo Martín. (2008). *Aplicacion de tecnicas de mineria de datos en la construccion y validacion de modelos predictivos y asociativos*. Salamanca.

Méndez, N. d. (2006). *Una Herramienta para la toma de decisiones*. Guatemala.

Montiel, G. G. (2006). *Auditoria Externa de servicios al cliente en una agencia automotriz*. Veracruz: Universidad Vercruzana.

Muñoz, H. R. (2002). *Diseño de un plan de implementación de data mining como soporte a la toma de decisiones con base en el registro académico de la Universidad Rafael Landívar*. Canada.

Noriega, R. M. (Diciembre de 2011). Agencia de Autos. 85.

Páez, J. A. (2003). *Avances en la toma de decisiones en proyectos de desarrollo de Software*. Sevilla.

Parsons, J. J. (2008). *Conceptos de Computacion, Nuevas perspectivas* (Decima ed.). Cengage Learning.

Peña Ayala, A. (2006). *Ingeniería de Software: Una guía para crear Sistemas de Informacion*. Recuperado el Diciembre de 2012, de http://www.wolnm.org/apa/articulos/Ingenieria_Software.pdf

Pérez López y Santin Gonzalez. (2007). *Minería de Datos, Técnicas y Herramientas*. Editorial.

Perversi, I. (2007). *Aplicacion de mineria de datos para la exploracion y deteccion de patrones delictivos en Argentina*. Argentina.

Pete Chapman, Julian Clinton, Randy Kerber and other. (1999-2000). CRISP-DM 1.0. En *Step-by-step data mining guide*. USA: SPSS Inc.

Piattini Velthuis, M. G., Marcos Martinez, E., Calero Muñoz, C., & Vela Sanchez, B. (2010). *Tecnología y diseño de Bases de Datos* (Primera Edicion ed.). Alfaomega Ra-Ma.

Pino, M. d. (Agosto de 2000). *El consorcio de CRISP-DM*. Recuperado el 29 de Noviembre de 2012, de Data Prix: <http://www.dataprix.com/prologo-crisp-dm>

Quispe Carita, V., Huamantuco Solorzano, D. H., & Vargas Yupanqui, J. L. (2011). Monografia Metodologia RUP (Rational Unified Process). *Tesis Monografica*. Puno, Peru.

Rao, G. y. (1996). *Data Warehousing, la integración de información para la mejor toma de decisiones*. Editorial Prentice Hall.

Ricardo, C. (2004). *Base de Datos*. McGraw-Hill.

Richard, D. (1997). *Organization Theory and Design* (6ta ed.). Estados Unidos : Internacional Thomson Publishing.

Rob, P., & Coronel, C. (2004). *Sistemas de Bases de Datos*. Thomson Learning.

Rodríguez, I. A. (2007). *Análisis de los factores que intervienen en la toma de decisiones de los administradores dentro de las organizaciones*. Queretaro.

Sumathi y Sivanandam. (2006). Introduction to Data Mining and its Applications. En *Studies in Computational Intelligence* (pág. 828). Alemania: Springer-Verlag.

Thearling, K. (2012). *Data Mining Techniques*. Recuperado el Noviembre de 2012, de <http://thearling.com/>

Trujillo, J. C., Mazón, J. N., & Pardillo, J. (2011). *Diseño y explotacion de almacenes de datos. Conceptos basicos de modelado multidimensional*. España: Club Universitario.

Universidad de la Salle. (Agosto de 2008). *Universidad de la Salle*. Recuperado el Agosto de 2012, de <http://unisalle.lasalle.edu.co/>

Universidad Yacambú. (s.f.). *Curso Teoría y Administración de Bases de Datos, UML*. Recuperado el Marzo de 2013, de http://www.oocities.org/es/avrrinf/tabd/Foro/Foro_UML.htm

Valda, J. C. (29 de Junio de 2009). *La toma de decisiones en la dirección de empresas*. Recuperado el 30 de Noviembre de 2012, de Grandes Pymes: <http://jcvalda.wordpress.com/2009/06/29/la-toma-de-decisiones-en-la-direccion-de-empresas/>

Vanrell, J. A. (2011). Un modelo de proceso para proyectos de explotacion de informacion. Mexico: Universidad Tecnológica Nacional (FRBA).

Witten y Frank. (2005). *Data Mining: Practical machine learning tools and techniques*. Estados Unidos: Morgan Kaufmann Series in Data Management Systems.

ANEXOS

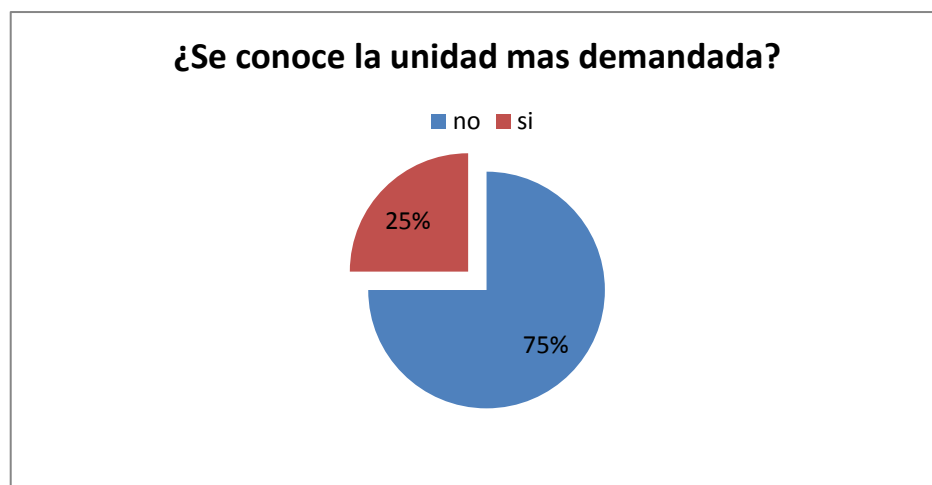
Anexo I. Cuestionario

Cuestionario para área de ventas de la agencia automotriz Renault.

1. ¿Consideras de importancia conocer en qué periodo las ventas han disminuido?
 - a. Si
 - b. No
2. ¿Le gustaría conocer el mes donde hay más movimiento de ventas?
 - a. Si
 - b. No
3. ¿Sería de apoyo para toma de decisiones contar con un comparativo de las ventas de diferentes periodos?
 - a. Si
 - b. No
 - c. Porque _____
4. Qué opina sobre la información con la que cuenta actualmente.
 - a. Es Completa
 - b. Incompleta
 - c. Otro _____
5. ¿Llevan un registro comparativo entre agencias?
 - a. Si
 - b. No
6. ¿Los clientes como se han enterado de las unidades con las que cuenta la agencia?
 - a. Redes Sociales
 - b. Publicidad
 - c. Familiares
 - d. Amigos
 - e. Otro _____

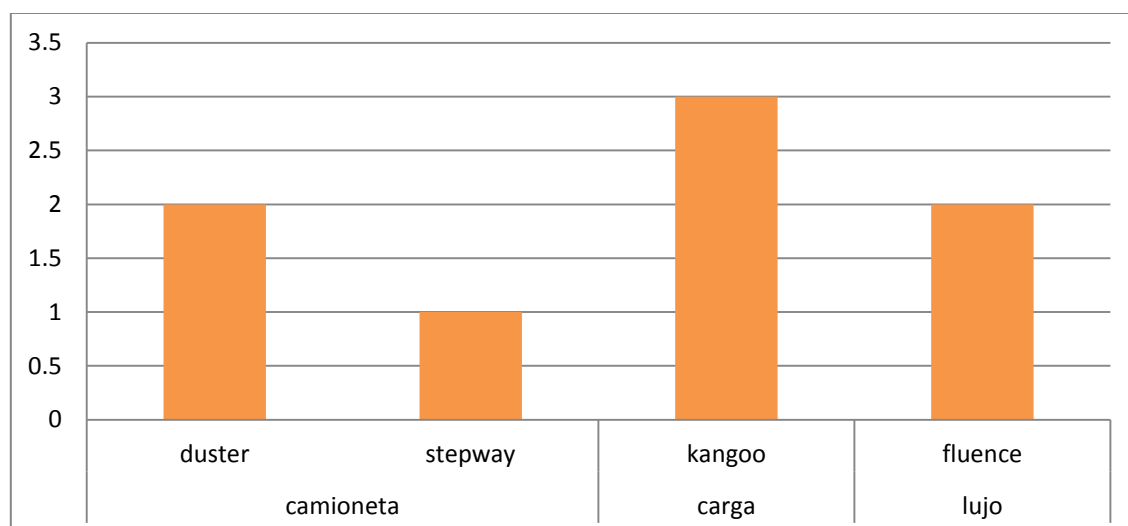
7. ¿Le gustaría conocer qué tipo de persona (física o moral) es la que más adquiere unidades?
- a. Si
 - b. No
8. ¿Qué tipo de persona física o moral actualmente es la que más adquiere unidades?
9. ¿Le gustaría saber que unidad es la más demandada?
- a. Si
 - b. No
10. Actualmente, ¿sabe que unidad es la más demandada?
- a. Si
 - b. No
11. Selecciona la unidad que consideras más vendida en los últimos meses:
- a. Duster
 - b. Koleos
 - c. Safrane
 - d. Kangoo
 - e. Stepway
 - f. Trafic
12. ¿Los clientes que tipo de unidades prefieren?
- a. Deportivo
 - b. Carga
 - c. Lujo
 - d. Camioneta

Los resultados que se muestran a continuación permiten determinar la situación actual del departamento de ventas, basados en el cuestionario aplicado e información que arrojo el sistema ERP, en la gráfica 1, se puede observar que el departamento involucrado desea saber que unidad es la más demandada para poder así tomar estrategias con las demás unidades.



Gráfica 1. Desconocimiento de unidades más vendidas
Fuente: Elaboración Propia

Así mismo, se observa en la gráfica 2, un comparativo de los diferentes tipos de unidades existentes según el modelo de unidad, donde usuario indica que unidad más solicitada es la Kangoo.

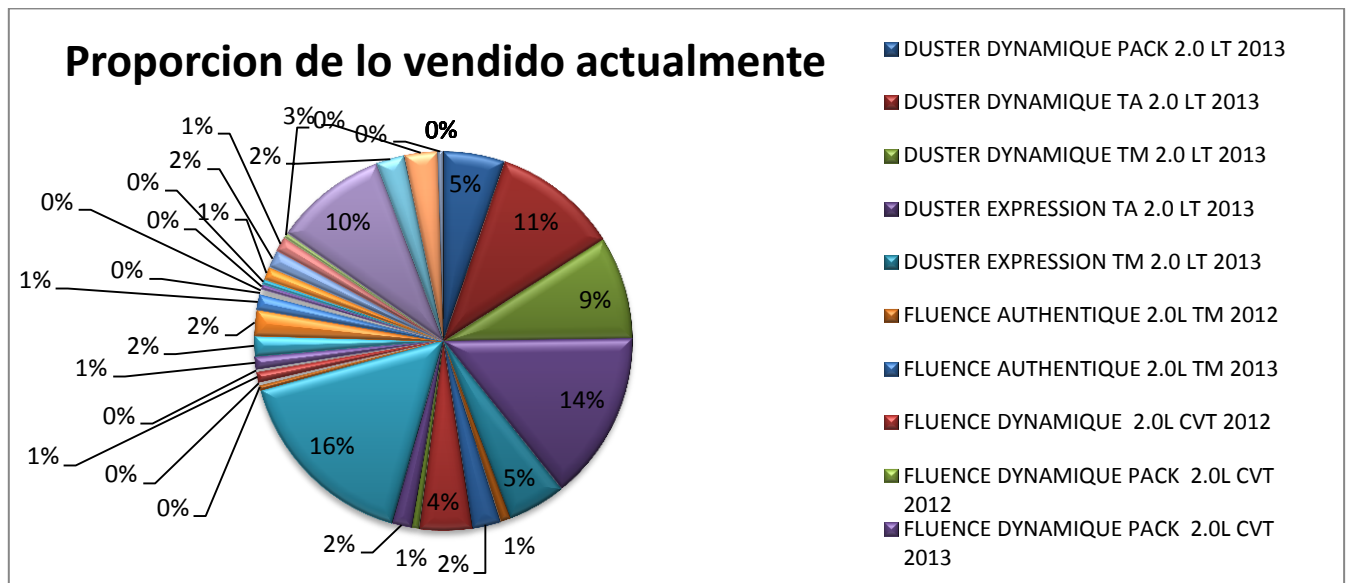


Gráfica 2. Tipo unidad según modelo
Fuente: Elaboración Propia

Proporcion de lo vendido actualmente

Modelo	Proporcion (%)
DUSTER...	23.00
DUSTER...	48.00
DUSTER...	39.00
DUSTER...	65.00
DUSTER...	22.00
FLUENCE...	4.00
FLUENCE...	10.00
FLUENCE...	19.00
FLUENCE...	2.00
FLUENCE...	8.00
FLUENCE...	75.00
FLUENCE...	2.00
FLUENCE...	4.00
FLUENCE PRIVILEGE...	4.00
KANGOO EXPRESS...	1.00
KANGOO EXPRESS...	5.00
KANGOO EXPRESS...	8.00
KOLEOS...	9.00
KOLEOS...	6.00
KOLEOS...	1.00
KOLEOS...	1.00
KOLEOS...	1.00
SANDERO 1.6...	2.00
SANDERO 1.6...	5.00
SANDERO 1.6...	7.00
SANDERO 1.6...	6.00
SANDERO 5...	2.00
SANDERO 5...	43.00
SCALA 4 PUERTAS...	10.00
SCALA 4 PUERTAS...	13.00
SCALA 4 PUERTAS...	2.00

La información obtenida se puede manipular y presentar de diferentes maneras, como se muestra en la gráfica 3, para conocer un porcentaje de las mismas ventas mencionadas en la gráfica 4.



[111]