

Inteligencia de negocios y minería de datos aplicado a la industria refresquera

Fani Rodríguez Flores^{1,2}, Leticia Flores Pulido¹, Emiliano Dávila de la Rosa²

¹Universidad Autónoma de Tlaxcala,
Departamento de Ciencias Básicas, Ingeniería y Tecnología,
México

²Ajemex, Big Cola, México

Resumen. El presente artículo sintetiza el desarrollo de una solución de minería de datos para la inteligencia de negocios aplicado al canal detalle, en una industria refresquera. El objetivo es desarrollar el análisis, diseño y construcción de una solución siguiendo los lineamientos de la metodología CRISP-DM. La solución de inteligencia de negocios que se utilizó, cuenta con un data mart con información consolidada, lo cual permitirá a los usuarios acceder más rápido a la información perteneciente a los indicadores comerciales, de esta forma los usuarios podrán tomar decisiones que concuerden con la realidad del entorno comercial de la empresa.

Palabras clave: Minería de datos, CRISP-DM, indicadores comerciales.

Business Intelligence and Data Mining Applied to the Refreshment Industry

Abstract: This paper summarizes the development of a data mining solution for business intelligence applied to the channel detail in a soft drink industry. The aim is to develop the analysis, design and construction of a solution based on the guidelines on the CRISP-DM methodology. The solution of business intelligence used has a data mart with consolidate information, which allows users to quickly access the corresponding commercial indicators information so the users can make decisions that match the reality of the business.

Keywords: Data mining, CRISP-DM, commercial indicators

1. Introducción

AJEMEX es una empresa refresquera, que pertenece al grupo de AJE es una empresa multinacional de bebidas, con presencia en más de 20 países de Asia, Latinoamérica y África, Cuenta con múltiples canales de distribución para llegar a distintos mercados o

segmentos. El Canal Detalle cuenta con una fuerza de ventas que se encarga de hacer contacto con los minoristas (detallistas), que venden los productos al público.

La empresa AJEMEX cuenta con diferentes Bases de Datos para poder almacenar su información, los datos de la preventa se almacenan en la base de datos de Aplicativo móvil mediante el dispositivo móvil, esta base de datos cuenta con más de 500,000 registros. La bases de datos donde se almacena la información procedente de los sistemas de Planeación de recursos empresariales (en inglés ERP, Enterprise Resource Planning) en la cual se realiza la facturación electrónica, esta base de datos cuenta con más de 3, 000,000 de registros.

Los usuarios necesitan de información segura y detallada, de cómo se realiza la atención al cliente en el canal detalle. Para realizar este análisis los usuarios solicitan al departamento de Tecnologías de la información la generación de diferentes reportes y tablas dinámicas, estas herramientas no proporcionan la facilidad de analizar la información, para poder ayudar a definir estrategia.

Los usuarios que toman decisiones en las ventas a detalle, se basan en los indicadores de mercado para poder definir estrategias de negocio para ciertos productos o fuerzas de ventas. La calidad, disponibilidad y presentación de la información juegan un papel decisivo en la toma de decisiones. Actualmente, se tiene un enfoque tradicional para el análisis de los datos donde se tienen fuentes de datos estructuradas y fuentes de datos no estructuradas.

Al aplicar técnicas de agrupamiento de minería de datos y análisis estadísticos para la obtención de información referente a los indicadores comerciales, es posible agilizar la definición veraz y oportuna de los indicadores para que sea más eficiente la toma de decisiones en el canal de venta a detalle, construyendo una Solución de Minería para Inteligencia de Negocios aplicado a canal detalle (SMINACD). El propósito es que la información relacionada con dichos indicadores pueda ser obtenida de manera eficiente veraz y oportuna. De esta forma, los usuarios podrán tomar decisiones que concuerden con la realidad del entorno comercial de la empresa.

2. Trabajos relacionados

2.1. Sistema de inteligencia de negocios para acueductos y alcantarillado

En el trabajo de Vanegas [1], los sistemas de procesamiento de transacciones (OLTP), dentro de los que se incluyen los sistemas de gestión, tienen diferentes limitaciones en cuanto a la forma de analizar la información y la usencia de reportes que muestren la información necesaria para la toma de decisiones. Mediante el uso de Inteligencia de mercados (BI) se logra unir el mundo de los datos y el de los negocios, esta permite a las empresas analizar grandes cantidades de datos de forma rápida y sencilla a la vez que facilita y apoya el proceso de toma de decisiones (PTD).

Se desarrolló e implemento un Sistema de Inteligencia de Negocios para las condiciones específicas de la EAALG (Empresa de Acueducto y Alcantarillado de la provincia Granma) permitirá mejorar la disponibilidad de información para el apoyo al PTD.

La metodología a utilizar es la HEFESTO de la cual se utilizaron las mejores prácticas, la metodología está compuesta por seis fases inicio, análisis de los

requerimientos, análisis de las fuentes de datos, modelado del DWH, integración de datos, representación de la información.

2.2. Minería de datos educativa

En el trabajo de Ballesteros [2], con el transcurso de los años, las actividades y el desarrollo de nuevas tecnologías se ha generado de forma considerable el almacenamiento de información, donde todo ese flujo de información que sea recolectado ha permitido satisfacer las necesidades diarias de las organizaciones, pero ha presentado un problema inherente en las capacidades humanas para analizar y transformar la información en conocimiento útil y relevante que apoye a la toma de decisiones, las técnicas y métodos de lo que se ha denominado actualmente como Minería de Datos (DM) enfocada a la gestión educativa, Minería de Datos Educativa (Por sus siglas en inglés EDM (Educational Data Mining), esta herramienta servirá a docentes que requieran evaluar sus prácticas y metodologías desarrolladas e implementadas con sus estudiantes, ya sea mediante el uso de agentes tutores inteligentes, sistemas virtuales de educación o estrategias activas de aprendizaje dentro del aula.

Se describe uno de los métodos analíticos para la selección de atributos y la extracción de parámetros de información KDD (Knowledge Data Discovery) conocido como el proceso de Descubrimiento de Conocimiento en Bases de Datos es un proceso no trivial de identificación de patrones válidos, novedosos y potencialmente útiles sobre un conjunto de datos, esto es, el objetivo es encontrar conocimiento útil, válido, relevante y nuevo sobre una determinada actividad, las etapas son Pre procesamiento, búsqueda (la utilización de técnicas y métodos de DM (Minería de datos)) y evaluación.

2.3. Un Modelo de Procesos para Proyectos de Explotación de Información

En el trabajo de Vanrell [3], dentro del desarrollo de proyectos de una empresa, los denominados proyectos de exploración de la información, poseen características propias que los hacen diferentes del resto, y no existe un modelo de procesos que se ajuste a este tipo de proyectos. COMPETISOFT es la proyección a nivel iberoamericano del modelo de procesos para el desarrollo de software MoProSoft, el modelo inicial fue modificado y adecuado a las necesidades de otros países, se le incorporo el modelo de evaluación EvalProSoft y se definieron niveles de madurez. La metodología CRISP-DM se encuentra definida en base a un modelo jerárquico de procesos.

El objetivo es crear un modelo de procesos de explotación de información orientado a Pymes tomando como base el modelo COMPETISOFT es la proyección a nivel iberoamericano del modelo de procesos para el desarrollo de software MoProSoft, esta metodología define un ciclo de vida de los proyectos de explotación de información que define las principales fases de un proyecto de este tipo. Estas fases son entendimiento de negocios, entendimiento de los datos, preparación de los datos, modelado, evaluación, despliegue.

3. CRISP-DM

La metodología de CRISP-DM para Minería de Datos [4], esta descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos. La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto como se muestra en la Fig. 1.

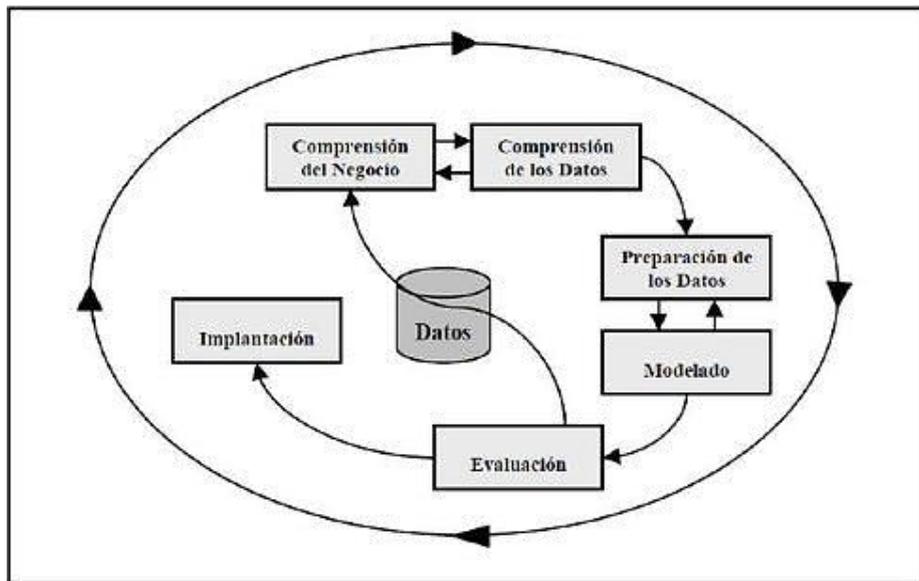


Fig. 1. Fases del proceso de modelado metodología CRISP-DM [4].

Comprensión del Negocio: La primera etapa de la metodología apunta a la comprensión de los objetivos del proyecto desde el punto de vista de los objetivos de negocio. En esta etapa se desarrollan los pasos preliminares para lograr los objetivos del negocio con herramientas de minería de datos.

Comprensión de Datos: En esta fase se desarrolla el entendimiento de datos y todas aquellas actividades relacionadas con la limpieza de datos, identificación de problemas vinculados con la toma de datos, procedimientos para determinar la calidad de datos y todo lo tendiente a facilitar la familiarización con los datos.

Preparación de los Datos: En esta etapa se desarrollan las actividades para construir el conjunto de datos final. Aquí se relaciona directamente el formato de los datos obtenidos con las herramientas de minería de datos a utilizar.

Modelado: En esta fase se eligen diferentes técnicas de modelado de datos y se estudian y ajustan parámetros con los valores correctos para el proyecto, Es altamente probable que desde el modelado sea necesario volver a la fase de preparación de datos puesto que todas las técnicas en evaluación pueden tener requisitos de formato de datos diferentes.

Evaluación: Hasta esta fase se han obtenido algunos modelos de minería de datos con sus datos y parámetros establecidos en forma óptima pero antes de pasar a la etapa final es necesario evaluar los resultados obtenidos por la ejecución de los programas en función de los objetivos de negocio. Aquí se puede presentar la necesidad de eliminar, modificar o considerar nuevas cuestiones relacionadas con el negocio.

Despliegue: El final del proyecto no termina con el modelado de datos y su ejecución y posterior evaluación de los resultados puesto que el conocimiento obtenido hasta aquí debe ser presentado de forma clara y precisa a todos los actores dentro de la organización, se puede presentar un simple informe de resultados, desarrollar una aplicación para la presentación de los resultados o bien instruir al usuario de los modelos para que ellos mismos generen y ejecuten los modelos con nuevos datos. Es importante al final de esta fase tener desarrollada toda la documentación del proyecto para dar independencia al usuario final en la utilización y generación de nuevos procesos de explotación de datos.

4. Minería de datos

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos [5]. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de los datos orientándolos hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining.

Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnología de reconocimientos de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos [5].

5. Inteligencia de negocios

El contexto de la sociedad de la información ha propiciado la necesidad de tener mejores, rápidos y más eficientes métodos para extraer y transformar los datos de una organización en información y distribuirla a lo largo de la cadena de valor [6]. La inteligencia de negocios (o Business Intelligence) responde a dicha necesidad, y podemos entender, que es una evolución de los sistemas de soporte a las decisiones (DSS Decissions Suport Systems). La inteligencia de Negocios busca transformar los datos en información para finalmente transformar la información en conocimientos. Algunas de las tecnologías que forman parte de Inteligencia de Negocios son data warehouse, reporting, análisis OLAP(On-Line Analytical Processing), análisis visual, análisis predictivo, cuadro de mando, cuadro de mando integral, minería de datos, gestión del rendimiento.

6. Aplicación de la metodología propuesta

Para el desarrollo de la solución de minería e inteligencia de negocios aplicado al canal detalle, se utilizó la metodología CRISP.DM, fue adaptada a las necesidades del proyecto. Las etapas que comprende la metodología CRISP-DM son las siguientes:

6.1. Comprensión del negocio

La primera etapa de la metodología apunta a la comprensión de los objetivos del proyecto los cuales son:

- Disponer de una base de datos que permita extraer conocimiento de la información histórica almacenada en la organización.

El contenido del data warehouse deben ser entendibles, navegables y su acceso debe estar caracterizado por el alto rendimiento. En esta etapa se obtuvieron los requerimientos de la solución como se muestra en la Tabla 1.

Tabla 1. Requerimientos de la solución.

Requerimiento	Necesidad
RF1	Obtener información en diferentes niveles de consulta
RF2	Obtención de indicadores comerciales: Cobertura, Efectividad de Visita, Drop Size, Liquidación, Efectividad de Compra.
RF3	Exportar información o cuadros de mando de los resultados.
RF4	Proveer una solución que permitirá a los usuarios finales generar sus propios reportes.
RF5	La solución deberá tener los mecanismos para controlar la seguridad de la Información.

7. Resultados de inteligencia de negocios

En la Empresa Ajemex se necesita conocer en qué zona y ruta se realiza mayor venta, para poder realizar la asignación de vendedores, correspondientes, se realiza en análisis de WEKA, contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización [10].

Para el desarrollo de esta solución se analizaran los datos provienen de la tabla ventas a nivel ruta por mes, la cual contienen 718,324 registros ,cuenta con una serie de atributos los cuales son número de región, código de la estructura, indicador, año, mes, tota, factor 1, factor 2, cuota por mes. Para el análisis de los datos se utiliza la técnica de árboles de decisión en la construcción del modelo a partir de los datos.

Un árbol de decisión es una estructura que permite dividir un extenso conjunto de datos relacionados entre s en conjuntos más pequeños de datos mediante la aplicación secuencial de sencillas reglas de decisión [7]. Se construir un primer clasificado para los datos, el algoritmo J48-C 25-M 2 de WEKA, El cual es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizado.

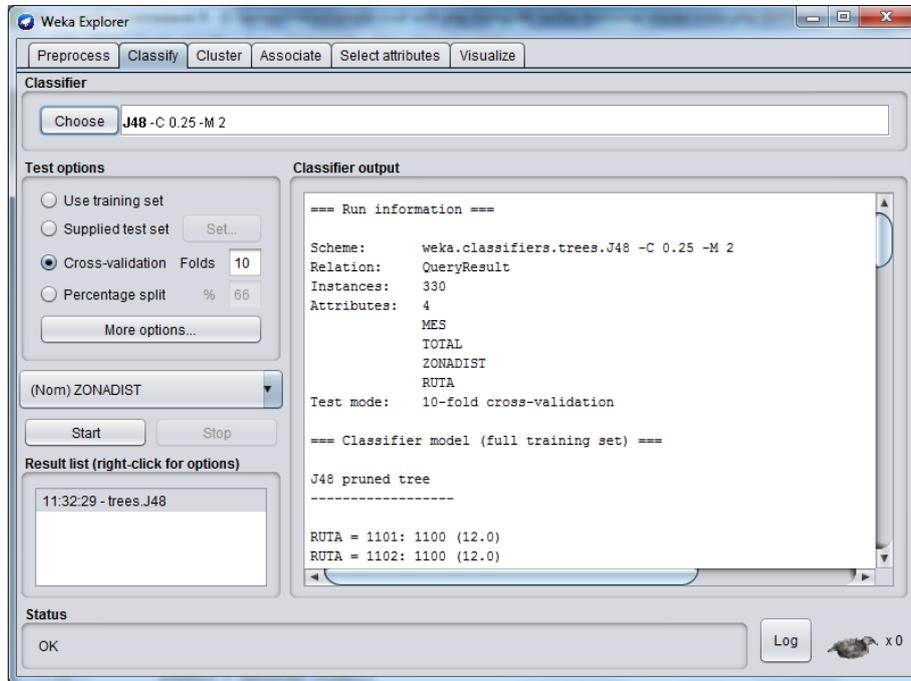


Fig. 2. Visualización para el análisis de datos por ZONADIST.

Para poder Realizar el análisis de los datos se realizaron pruebas por Zona de distribución, ruta, Mes. En la Fig.2, muestra información sobre el tipo de tipo de clasificador utilizado, el archivo el cual se está trabajando, el número de instancias (330), el número de atributos 4, y sus respectivos nombres.

En la Fig. 3. se muestra la matriz de confusión por Zona de Distribución (ZONADIST), la cual podemos observar que los valores de la diagonal son los aciertos y el resto lo errores, podemos observar que la zona 1100, se clasifican correctamente 89.

```

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  <-- classified as
89  0  0  0  0  0  0  0  | a = 1100
 0 28  0  0  0  0  0  0  | b = 1200
 0  0 35  0  0  0  0  0  | c = 1300
 2  0  0 62  0  0  0  0  | d = 1400
 0  0  0  0 24  0  0  0  | e = 5000
 0  0  0  0  0 40  0  0  | f = 1500
 0  0  0  0  0  0 30  0  | g = 1600
 0  0  0  0  0  0  0 20  | h = 1700
    
```

Fig. 3. Matriz de confusión por ZONADIST.

En la Tabla 2 se muestra una comparación del porcentaje de clasificación de las instancias, en la cual podemos observar que la clasificación por ZONADIST, nos proporciona un mejor porcentaje de instancias correctas.

Tabla 2. Tabla de hechos.

Nombre	Necesidad	Número de Registros
VTASRUTAM	Ventas a nivel ruta por mes	71,823
VTASRURAD	Ventas a nivel ruta por d a	127,4761
VTASCEDID	Ventas a nivel cedi por d a	44,471
VTASCEDIM	Ventas a nivel cedi por mes	1,885

7.1. Herramienta de inteligencia de negocio PowerPivot

PowerPivot es una verdadera revolución dentro del análisis de datos porque nos da todo el poder que se necesita para realizar complejos análisis sin requerir la intervención de técnicos. Esta herramienta, un complemento Excel, que contiene una base de datos tipo OLAP en una memoria interna [8].

Para que los usuarios finales puedan tomar decisiones que concuerden con la realidad de la empresa, utilizamos la herramienta PowerPivot la cual transforma una solución de inteligencia de negocio impulsada por usuarios, sin depender del departamento de TI, para la generación de tablas dinámicas, cuadros de mando, reportes, etc. Al seleccionar las tablas de hechos y dimensiones, enseguida se despliega los datos de las tablas seleccionadas. PowerPivot realiza una copia de los datos en Excel. Millones de datos se pueden almacenar de manera efectiva por PowerPivot utilizando una nueva tecnología que comprime los datos. Esta herramienta nos permite exportar a Excel, la información de diferentes maneras, en la Fig 4, se muestra un cuadro de mando, el cual nos muestra información de los indicadores comerciales, la información cambia en base al nombre de la sucursal, mes y año.

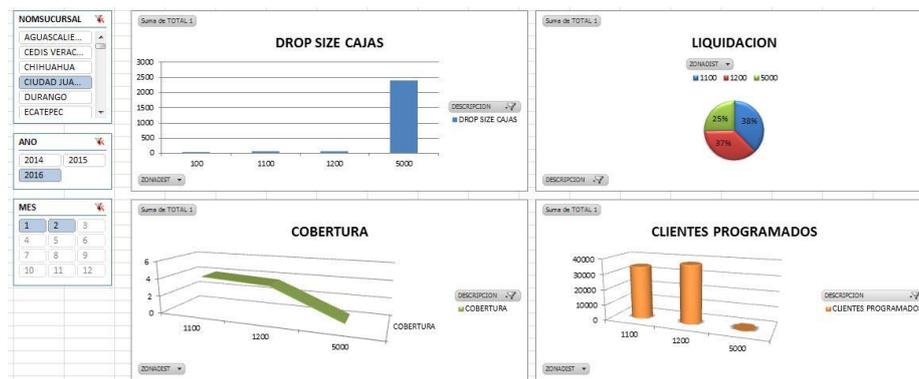


Fig. 4. Cuadro de mando con los indicadores comerciales.

8. Conclusiones y trabajos futuros

Hoy las empresas manejan gran cantidad de datos, se necesitan diferentes herramientas para poder convertir los datos en información, la cual el usuario pueda explotar y analizar. Para poder tomar decisiones se desarrolló una solución para inteligencia de negocios aplicado a canal detalle, para la empresa refresquera Ajemex, se realizó un análisis de la información referente a los indicadores comerciales, mediante la metodología CRISP-DM. Al término del análisis se obtuvo como resultado un Data Marts de ventas, el cual obtiene la información de una manera más rápida. Se utilizaron técnicas de minería de datos, para poder clasificar la información, tales como arboles de decisión (algoritmo J48), se utilizaron los métodos para poder obtener información referente a las ventas por zona.

La Metodología CRISP-DM, facilitó el poder realizar un proyecto de exploración de la información, al utilizar herramientas como PowerPivot se le proporciona al usuario la capacidad de poder realizar su análisis y reportes sin necesidad de depender del área de TI, esta herramienta es fácil de utilizar ya que los usuarios están familiarizados con Excel, solo será un complemento para poder realizar sus diferentes reportes. En trabajos futuros se puede utilizar el algoritmo analizado, para poder integrarlo al PowerPivot y nos pueda ayudar a predecir como será la venta en diferentes rutas.

Referencias

1. Vanegas, E.: Sistema De Inteligencia de Negocios Para Acueductos y Alcantarillado. 3C TIC, 7ma Edición (2013)
2. Ballesteros, A.: Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. Lat. Am. J. Phys. Educ., Vol.7
3. Vanrell, J. A.: Un Modelo de Procesos para Proyectos de Explotación de Información. Lacrete, Medellín (2012)
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 step-by-step data mining guide. Technical report (2000)
5. Pérez, C.: Minería de Datos, Técnicas y Herramientas. Thomson Ediciones Paraninfo, S.A., 1ra Edición (2007)
6. Conesa, J.: Introducción al Business Intelligence. Editorial UOC, 1ra Edición (2010)
7. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann. United States of America (2016)
8. Muni, L.: PowerPivot con Excel a su alcance para convertir sus datos en información. Pro t Editorial I., S.L., Barcelona (2012)
9. Portada sobre la plataforma Pentaho Open Source Business Intelligence, En línea, <http://pentaho.almacen-datos.com/kettle-spoon.html>
10. Manual de WEKA. En línea, <http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/weka.pdf>